13

Evaluation Studies

Stephen D. Lapan

Northern Arizona University

Forthcoming in:
Foundations of Research: Methods of Inquiry in Education and the Social Sciences

Edited by
Kathleen Bennett deMarrais
The University of Georgia
and
Stephen D. Lapan
Northern Arizona University
Lawrence Erlbaum Associates

Meet the Author

Stephen D. Lapan earned an MA at the University of Illinois, a Ph.D. in Educational Psychology at The University of Connecticut, and is currently a professor in the Center for Excellence in Education at Northern Arizona University (NAU). He is director of the doctoral program in Curriculum and Instruction and teaches courses in approaches to research and evaluation. He is Consulting Editor for the Journal of Research in Childhood Education, past Editor of the Excellence in Teaching Journal, and past Editor of the Center for Excellence Monograph Series.

Dr. Lapan's selected publications include a book- *Survival in the Classroom* (1978) (with E. House); book chapters- *The Evaluation of Teaching* (1989) and *Policy, Productivity, and Teacher Evaluation* (1997) (with E. House); and journal articles- *Guidelines for Developing and Evaluating Gifted Programs* (1989), *Evaluation of Programs for Disadvantaged Gifted Students* (1994) (with E. House), *Criteria for Gifted Programs (1996), and Students as Independent Learners* (1998) (with P. Hays).

Dr. Lapan's awards include NAU charter Faculty Fellow (1988), first NAU Teaching Scholar (1989), and the Arizona Association for Gifted and Talented Life Achievement Award "for service to gifted education." His selected studies include evaluation of the Illinois Gifted Program for the Illinois legislature (1968-1970), assessment of the Florida Accountability Program for the National Education Association (1978), evaluation of the Cooke Magnet School for Gifted Children (1981), and evaluation the NAU Jacob Javits Getting Gifted Grant (1993).

Background[1]

The field of professional evaluation has developed over the past three decades as an interdisciplinary field with researchers from many areas of social science contributing to its concepts and practice. There are professional evaluation associations in more than twenty countries and half a dozen journals that specialize in evaluation. Evaluation studies are typically sponsored and funded by branches of government, including federal, state, and local agencies. Studies are carried out by professors, independent consultants, or companies that specialize in such work.

The accepted definition of evaluation is the determination of the merit or worth of some entity. The basic difference between evaluation and other forms of social research is that evaluators arrive at conclusions such as, "X is a good program or has merit," while other researchers arrive at conclusions such as, "X causes Y." Some overlap can be found between evaluation and other forms of research, however, where evaluators sometimes look for causes as well as quality and other researchers might produce conclusions about program quality along with making cause and effect connections.

The logic of evaluation consists of four steps:

1. Establishing criteria of merit
2. Constructing standards
3. Measuring performance and comparing it to standards
4. Synthesizing and integrating the performance data into conclusions of merit and worth

This reasoning is familiar from publications like Consumer Reports. In evaluating cars, evaluators decide which specific cars to evaluate. Then they establish criteria of merit, such as acceleration, handling, durability, and cost. They measure the performance of the cars and compare it to the standards they have adopted. Sometimes the standards are comparisons with the other cars, such as "This car is best at handling." Other standards may be derived from expectations people have about cars. "This car turns over in emergency maneuvers."

These performance results are merged into conclusions about which cars are best among those studied or which cars are acceptable. Of course, it is the case that the cars differ in performance on different criteria. The car that handles best may not be the cheapest. Evaluators must somehow synthesize these performance results into summary conclusions, leaving consumers to make final decisions.

What is evaluation and what does it accomplish? In what contexts are its applications meaningful and needed?

Evaluating educational programs is different in some ways. Programs serve constituencies and what is best for one constituency on one evaluation criterion may not be best for others. Evaluators must weigh and balance these different considerations. Again, the evaluation criteria are derived from the nature of the entity being evaluated and what people expect from it. For example, what do people expect from reading programs: that the students improve their reading speed or analytical skills, improve their reading for content or enjoyment? Evaluation audiences are likely to identify many different criteria for success in reading depending on what they value. Deriving criteria and standards of performance are critical components of evaluations. Once the criteria

[1] Ernest R. House served as contributing editor for this chapter.

are derived, which may take some effort, evaluators use the data collection methods of social research to measure performance.

The focus of evaluation can be **programs**, **personnel**, **products**, **materials**, and **policies** though the focus for this chapter is on programs. **Personnel evaluation** entails observing an individual's performance and what the person produces. For example, professors are evaluated on the basis of their teaching and their scholarly publications. Classroom teachers are judged on their ability to teach effectively and on student achievement. Evaluating **products** and **material** entails determining whether they meet the purposes set out for them. For instance, instructional materials in mathematics might be evaluated according to the standards of the National Council of Teachers of Mathematics (NCTM) and whether teachers and students can effectively apply these materials in classroom work.

Policy evaluation focuses on the effects of policies established by government agencies. A policy evaluation might ascertain the effects of retaining students by determining whether the failed students achieve more, drop out of school later, or suffer emotional problems from failing a grade.

Finally, **program evaluation** emphasizes how educational and social programs are implemented, how they operate, and what effects they have. In a school setting, a sixth grade science program might be evaluated by studying how the program began, what happens in class, and what effects the program has on students, teachers, and parents. Evaluators might examine written program descriptions and talk with planners about the history of the program. Evaluators also might observe classes, interview students, teachers, administrators, and parents.

Fundamentally, there are about half a dozen different approaches to program evaluation. One of the most common approaches is to determine if the program is achieving the goals set out for it. If the program is designed to reduce school dropouts, does it accomplish this? If the goal of the program is to increase female participation in math and science classes, does it do this?

There are other ways that programs are evaluated too, including the use of testing, indicators, and experiments. The use of standardized tests dates back many decades. Although standardized tests were originally developed to ascertain the educational needs of individual students (Binet & Simon, 1905), they currently are a dominant method for assessing educational programs. Tests are efficient to administer, require no outside expertise to analyze, and meet the reporting demands made by local school boards and state and federal agencies.

One task of program evaluation should be to establish a clear connection between the program and its effects on students. Tests do not always allow these connections to be made clearly because test scores are influenced by many factors including student ability, language background, and socioeconomic status. Thus, changes in test scores might be caused by factors that have little to do with the quality of the program being evaluated. While tests can make a contribution to an evaluation, their limitations make them unsuitable as the sole means of measuring program effectiveness.

Measurement of indicators is another method used to determine program success. Some indicators can be tied to program goals and can be sensitive to the particular program. Indicators might include parent requests for student inclusion, student dropout rates, teaching effectiveness, student classroom performance, and student ratings of the program. Using several indicators rather than one provides a more comprehensive picture of the program's influence. A potential shortcoming of this indicator approach is failure to discover unknown effects. If only program goals are used to generate indicators, unanticipated program effects might be overlooked.

Many experts have called for the inclusion of experimental methods to increase the rigor of an evaluation. Comparison and control groups permit more precise estimates of program effects in some cases. Well-conducted experiments rule out rival explanations for effects, such as student ability or socioeconomic status, and provide strong evidence. Indeed, experiments can work well in situations where the rules of rigorous investigation can be followed. However, in settings like schools, experiments are difficult to use since experimental manipulation and random assignment of students, teachers, and treatments (usually instructional programs) are not ordinarily possible. Parents do not like their children to be randomly placed in classes. When experiments are conducted, considerable expertise is also required in carrying out and interpreting results correctly.

What are the pitfalls of using any single approach to evaluation such as just goals or just test scores?

Qualitative Versus Quantitative Evaluation

In past years, there was a debate over whether evaluations should be based on qualitative or quantitative data. Qualitative proponents argued that thick descriptions and particular knowledge gained from program participants outweighed quantitative indicators like test scores. Quantitative proponents argued that test scores and other numerical findings provided more objective evidence of the effects of programs. This debate has been resolved by both sides recognizing the place of both quantitative and qualitative data and that the best studies would incorporate both kinds. However, selecting data collection methods should not be the first concern in planning evaluations. Evaluations should be based on the content, purpose, and outcomes of the program, rather than being driven by data collection methodologies.

Doing Program Evaluation

For convenience, school settings have been used in this chapter to illustrate program evaluation. However, just as evaluators can evaluate school districts, schools, and classrooms, they also can carry out evaluations of counseling activities, social service agencies, and federal and state educational delivery efforts.

First Steps

Program evaluations typically begin when program developers or school leaders contact a professionally trained evaluator (often located at a university), requesting an evaluation of a program. The evaluator meets with those who want to sponsor the study to discuss why the evaluation is being done and for whom. The "why" answer might range from local concern over improving a program to reporting requirements of funding agencies. The "for whom" answer relates to who has a stake in the program and who wants to know how well the program is working. It is important for evaluators to find out the limitations placed on the study and who the audiences are.

At this point, the evaluator develops a description of the program based on information obtained from early meetings and from documents such as program proposals. The evaluator might talk with teachers and other participants who have not been in the meetings to be certain that all appropriate perspectives are represented in the plan. This program description might be shared with sponsors and program participants to gain clarity about how everyone sees the program at this stage. This exchange encourages different views of the program to be expressed and promotes a broader understanding. During these meetings, the evaluator might request that sponsors and participants indicate evaluation questions they want answered. Other questions may come from outsiders, such as school board members, those who fund the program, or program evaluators themselves.

It is the evaluator's job to develop a plan based on program descriptions, needs of the stakeholders and audiences, and the specific questions generated, as well as raise issues and questions not covered by those interviewed. The plan could include the program description along with evaluation questions and a time line for collecting data and reporting findings. Some plans are detailed in specifying exactly how questions will be answered. Other plans are less specific, perhaps listing a few evaluation questions and deadlines for the work.

Evaluators address several questions in each evaluation. Ordinarily, they also use multiple sources of data and multiple methods of collecting data to answer the questions. In addition, the sponsors of the evaluation and the evaluator should develop an evaluation agreement between them to clarify these issues. This agreement includes the distribution of responsibilities for the evaluation and who will see the results of the completed study. The agreement answers questions as to who will conduct the evaluation, who will write and present the findings, when the evaluation will be completed, in what form the results will be reported and to whom, whether preliminary findings will be reported early for purposes of feedback, whether there will be a minority report if there are disagreements with findings, and the costs of the study. For example, sponsors may not want parents to know that the program is not working if it is not. If parents are stipulated as audiences in the agreement, chances are reduced that they will be left out. Putting such issues in the agreement reduces conflict later.

How is it important that evaluation, like most research, should be designed around content and questions rather than techniques for collecting data?

Data Collection

Instrument selection and development are key components in planning. The quality of the data is only as good as the instruments used to collect it. Ordinarily, evaluators develop new items or protocols since the purposes of programs vary and often there are no standardized instruments to fit the particular characteristics of a given program. Whether adaptations of existing instruments are made or new ones are developed, testing the instruments in the field is essential to their readiness for use. This field testing also provides practice for the novice evaluator in administering questionnaires, conducting interviews, and observing classrooms.

In evaluation studies, especially those using qualitative data, the evaluator is an integral part of the data collection. This is particularly so in observation and interview work where the evaluator's judgment and influence on those being observed can influence the data collected. For example, in an interview the evaluator might ask questions that encourage frank responses, thus making the evaluation results trustworthy. Or, the evaluator may convey that certain answers are desirable, resulting in information influenced as much by what the evaluator believes as what the interviewee was thinking.

The major data collection techniques used in program evaluation are interviews, observations, tests, questionnaires, and inspection of documents.

Interviews. The face-to-face interview is one of the best sources of information. Perspectives gained through this give-and-take process represent more than points of view; they offer insights into special knowledge that only participants possess. For example, teachers can recount how classes work or what is emphasized in lessons. Students can disclose personal views of what class is like for them. Interviews sacrifice coverage to gain depth, but depth is important although time-consuming.

Developing and conducting interviews are not easy tasks. Interview protocols developed by professionals can be used, but most must be adapted for particular programs. In developing interview questions, one must avoid questions that suggest certain answers (e.g., The parents really like the program, don't they? Would you say the materials are the real strength of the program?). Novice evaluators need to practice using protocols and listening to their taped interviews to sharpen their questioning ability.

Observations. Observations are used to gain insights about program operations. While there are many observation schemes available, the evaluator may have to adapt or develop one that suits the program. Again, practice and study is necessary to learn effective observation techniques. For instance, it may be helpful if two observers watch the same taped lesson independently and compare how each described what was happening. A few practice attempts improve the accuracy of the observations.

Tests. The use of standardized tests is a popular method for data collection in program evaluation, but such scores do not adequately reflect the content and purposes of any given program under review. A better choice would be to construct tests that parallel the students' program learning experiences, although test reliability and validity evidence should be produced for these measures. As noted earlier in the chapter, tests of any kind should not be the exclusive means for evaluating a program.

Questionnaires and document inspection. These approaches are useful in providing background information. Questionnaires might be sent to parents to verify their child's involvement in a program and obtain their views on observed effects at home. Documents may reveal aspects of program history and goals that can be corroborated with sponsor and participant views. These approaches do not explore the depth of meaning that interviews do, but can complete the picture the evaluator is developing.

Sampling

Purposeful sampling, the deliberate selection of information-rich sources, drives most program evaluation efforts. This sampling procedure is part of the evaluation plan where the best data sources are defined by the study questions. Program teachers know most about day-to-day lesson planning while administrators know most about program funding. Additional sampling occurs during the evaluation when evaluators learn of information sources they could not have known about when the evaluation was planned. This is known as "emerging design," adjusting evaluation plans as the study unfolds.

Evaluators sometimes use random sampling to obtain statistically representative samples so that generalization to larger populations is accurate. This sampling approach is used in cases where large groups are being surveyed. For example, analyzing test scores of large groups of students in school districts requires knowing precisely which students the data came from or what groups were represented in the data collected.

Trusting Evaluation Results

The term "validity" is commonly used in research to describe whether a study's results can be believed. Trusting results in program evaluation is a serious issue especially when qualitative data are employed. Sometimes audiences are unfamiliar with this kind of information and may want numbers and percentages so they can feel more secure about the objectivity of the findings. Two approaches used in validating qualitative studies are **member checking** and **triangulation**.

Member checking. Program evaluation often involves many program participants over extended periods of time. Data are collected at several intervals, and the evaluator may choose to provide preliminary summaries to participants. Unlike traditional research where data are gathered and analyzed before reporting, in program evaluation it is appropriate to share findings during the study, especially when participants have finished providing observations about the program. By sharing preliminary findings, the evaluator is able to gauge how early results fit with the understanding of participants and sponsors. This participant or member checking allows teachers and others to question the findings or request clarification, thus challenging the evaluator to reveal evidence, change interpretations, or collect additional data. Sharing increases evaluator credibility.

Triangulation. Triangulation, another technique, refers to the collection of data from two or more sources (e.g., students, teachers) using two or more methods (e.g., interviews, observations). For example, if an evaluator wants to learn how the program operates at the classroom level, use of classroom observations (a method), student interviews (a source and a method), and teacher interviews (different source, same method) produce the triangulation. Overlapping evidence is generated rather than obtaining just one perspective. Each method and source has strengths and weaknesses, and using several methods and sources builds on strengths.

Another form of triangulation has been found useful when the evaluator is also a program participant and is possibly biased by being too close to the program. Researcher triangulation is where an outside evaluator analyzes the data collected by the in-house evaluator and draws conclusions without knowing the insider's interpretation. These conclusions are compared.

How can evaluation studies be considered valid? When would evaluation results not be trusted?

Analyzing and Reporting Study Results

Evaluation studies transform interview, observation, test, questionnaire, and document data into descriptions and explanations. Often these data are voluminous and must be interpreted by reading over the material several times. The evaluator might begin by highlighting thoughts or phrases that emerge as patterns, then writing summaries that capture these emerging patterns. In most cases, these summaries are formed by linking the findings with the evaluation questions formulated at the outset. This approach allows the evaluator to discover effects not anticipated in the original evaluation plan, while adhering to the original evaluation questions.

These findings can be presented in writing as well as through videos, multimedia presentations, and discussion formats. The goal is to formulate the findings so that audiences understand and respond. For example, educators and parents need different words and images to comprehend meaning. It is also helpful that findings are linked to recommendations that give stakeholders guidelines to follow for improvement, though making recommendations is not a necessary part of the evaluation. This depends on what was agreed to at the beginning of the study.

Final evaluation reports contain complete findings and data summaries where possible. Minority reports and testimony from those who disagree with the findings can make the reports more balanced and credible. Also, frequency counts and percentages improve clarity, but should not replace description and rich detail. Final reports usually begin with an executive summary containing concise statements of the evaluation questions, findings, and recommendations (if requested). A rule of thumb is that a summary longer than three pages exceeds the attention span of audiences. The length of the total report depends on the breadth and depth of the evaluation.

An Example

A practical example will provide a more concrete way of thinking about how evaluations work. What follows is the description of an educational program and how it might be evaluated.

Program description. An independent study program has been developed for seventh graders in two social studies classes for those students who complete required assignments early or demonstrate a grasp of material before it is introduced. In most cases, the two classroom teachers use pre-tests and/or checklists to determine which students might pursue independent projects, but students also may volunteer for project work if the teachers agree.

The two primary goals for the Independent Study (IS) program were to offer students opportunities to gain greater depth in school curriculum areas and to learn how to become more self-directed learners. Consequently, selected IS students were encouraged to study areas that are linked to classroom content and challenged to take more responsibility for their own learning decisions.

Typically, selected IS students develop a contract with their teacher that specifies project content, deadlines, proposed products, and plans for evaluating the end results. Students would devise project schedules that could include all or a portion of their social studies class on a daily, weekly, or monthly basis. Teachers would hold planning, progress, and culminating conferences with each student throughout the IS schedule. Planning would include topic selection, setting deadlines, determining work locations and needed resources, and devising plans for judging project results. Since IS students receive grades for their required classroom work in social studies, no grades are given in IS. This is intended to encourage more student responsibility over decisions about learning rather than waiting for teacher direction. Progress conferences are used to monitor student work while culminating meetings address project results, effectiveness of student self-direction, and any future project plans.

The IS program has been in operation for a little more than two school years and approximately 15 students from the two classes have been involved in the IS project work each year. These students have produced video tapes, articles, small books, slide and PowerPoint presentations, and three-act plays on topics such as the World War II

Japanese-American concentration camps, power and corruption, causes of war, and gender politics in the 20th Century.

The teachers have concluded that an evaluation would help them decide if the program should be retained or at least improved. They believe it is working well, but know that their judgments may be clouded by closeness to and personal investment in the program. A professor from the local university has agreed to conduct the evaluation. She is an experienced evaluator who has studied similar IS programs in other schools.

Evaluating the IS program. Following the framework outlined in the chapter, the evaluator would seek written and verbal descriptions of the program to obtain a preliminary idea about its purposes and functions. Soon after this, she would want to determine the evaluation's focus and specific questions by talking with the program teachers, some program students, and any administrator who may know about the IS program.

In this instance, the program evaluator used input from these individuals along with her own expertise to develop a list of areas the program teachers supported. These areas usually explicitly or implicitly offer the criteria and standards crucial to conducting any evaluation, but often are not labeled as such. These areas would then be used as guides for the evaluator in developing questions to be answered in the study. In this evaluation study example, the areas developed were:

- Appropriateness of student selection
- Teachers' role in encouraging/discouraging independence
- Students' role in encouraging/discouraging independence
- Resources and environments that encourage/discourage independence
- Student progress toward more self-directed learning
- Student depth of understanding in the curriculum
- Quality of student independent study products
- Continuity of monitoring student independence from one project to the next

In addition to this list, the teachers wanted to know about the program's overall strengths and weaknesses and what they might do to improve things where needed. Subsequent to this planning, the evaluator would develop evaluation questions reflecting the areas of concern and ways to answer these. Note that the areas as well as the evaluation questions that follow below go well beyond just determining if the two program goals are met (content depth and self-direction).

First, the evaluator may want to determine how students are selected to participate in the IS program. Do students have an equal opportunity if they meet the stated criteria? Are teachers making selections based on reasons other than those indicated? Why do some students who volunteer become IS students while other student volunteers do not? Here the evaluator would piece together the selection process by interviewing teachers and students and by examining the match between classroom content and the pre-test and/or checklist items. Sifting through these data will allow the evaluator to construct how the selection process works along with ways it might be improved.

Following this, the evaluator will want to study how the program operates once students are identified. Of particular interest here and elsewhere will be the extent to which students are encouraged to be self-directed (one of the two primary program goals). Questions to be answered at this stage might include:

1. What occurs during the student-teacher planning conferences?

2. How are topics and projects selected and by whom?

3. How are study schedules, plans, work locations, and resources decided and by whom?

4. To what extent are student independence and decision making encouraged or discouraged by the teachers?

5. Overall, what roles in decisions are taken by teachers and by students?

Again, following the guidelines presented in the chapter, multiple methods and sources would be used to increase the validity of the findings. The evaluator would probably observe and/or tape record planning conferences as well as interview students and teachers about what usually occurs during these meetings. The evaluator would use this information to develop answers to the questions posed, possibly uncover other important findings not anticipated, and provide an explanation of how the planning conferences operate.

It also would be useful to study how the program works for students once they are engaged in independent study, once again by observing conferences and interviewing teachers and students. In addition, observing students while they work on their projects may provide useful insights. Are students working alone? Do they have access to necessary resources? Does their workplace support their project efforts? Are projects linked to the regular curriculum (the other primary program goal)? The answers to such questions will provide the evaluator with case scenarios that typify different project work patterns and might reveal needed changes in how students are guided or supported in these efforts. Also of interest would be the determination of how students are handling the new learning environment. It is often at this stage, for example, that students feel lonely or lost since they are not able to gauge their progress compared to their peers. Perhaps, on the other hand, the teachers have prepared the students so that these concerns do not arise.

Another step in the evaluation process might be to observe culminating conferences as well as inspect the projects developed by the students. In addition, it would be useful to find out how the IS students view the experience and whether or not they are gaining confidence in making more independent decisions. Questions in this portion of the evaluation could include:

1. In which areas are students making study decisions and in which areas are teachers taking the lead?

2. What are students learning beyond what they have obtained from regular class material?

3. What is the quality of the final IS projects?

4. How are the projects evaluated and by whom?

5. Is curriculum content or becoming more independent emphasized most in the culminating conferences?

6. To what extent are future projects and ideas about becoming more self-directed emphasized during these culminating conferences?

7. Why are some IS experiences effective while others are not?

It may appear that there is no end to the areas and questions that might focus an evaluation. Important guiding factors, however, limit the evaluator's scope. The needs of the primary audiences and stakeholders are paramount in designing the evaluation as well as the expert judgment of the evaluator. Limits are in place as well. The amount of time and resources often constrain what the evaluator can accomplish.

In this plan, administrators and outside agencies were not the primary audiences while the teachers and students were. And, time and limited resources may not allow the evaluation to include parent perspectives, for example, although such data are often useful. Thus, while there is no one way to design and conduct an evaluation, selection of methods and data sources are formulated with the stakeholders and limitations in mind.

What is the significance of audiences and stakeholders in evaluation studies?

Acknowledgement: The author wishes to thank those who provided insightful comments on earlier drafts of this chapter. They are: Deanne Falls, Gila Garaway, Patricia Hays, Bob Holloway, Dan Kain, Ellen Menaker, Frances Riemer, Enid Rossi, and Frank Vogel.

Evaluation Resources

Books

House, E. R., & Howe, K. R. (1999). *Values in evaluation and social research*. Thousand Oaks, CA: Sage.
An advanced analysis of the value issues in evaluation and how they might be resolved.

Madaus, G., Kellaghan, T., & Stufflebeam, D. L. (Eds.). (2000). *Evaluation models*. Boston: Kluwer Academic Publishers.
An analysis of the different approaches to evaluation in the language of the originators.

Patton, M. Q. (2002). *Qualitative research and evaluation methods* (3rd ed.). Thousand Oaks, CA: Sage.
A very comprehensive and easily accessible introduction to evaluation methodology and issues.

Shadish, W., Cook, T., & Leviton, L. (1995). *Foundations of program evaluation*. Thousand Oaks, CA: Sage.
A more advanced analysis of the views of the founders of the field.

Worthen, B., Sanders, J. R., & Fitzpatrick, J. L. (1997). *Educational evaluation* (2nd ed.). Reading, MA: Addison-Wesley.
A textbook on various approaches to evaluations with detailed examples.

Associations

American Evaluation Association (AEA) http://www.eval.org

American Educational Research Association (AERA) http://www.aera.net/

Journals

American Journal of Evaluation (formerly *Evaluation Practice*) published by AEA

New Directions for Evaluation published by AEA

Educational Evaluation and Policy Analysis published by AERA

Evaluation published by Sage in London and Thousand Oaks, California in association with The Tavistock Institute (London, UK)