Evaluation Domains: Program Evaluation

Christina A. Christie

tina.christie@cgu.edu

Claremont Graduate University

Leslie A. Fierro

Leslie.Fierro@cgu.edu

Claremont Graduate University
Name & affiliation as it should appear in the finished work:

Primary author:
Name:
Christina A. Christie, Ph.D.
Affiliation:
Claremont Graduate University, School of Behavioral and Organizational Sciences
Full Address (for proofs only):
370 Cherry Drive
Pasadena, CA 91105

Co-author:
Name:
Leslie A. Fierro, MPH
Affiliation:
Claremont Graduate University, School of Behavioral and Organizational Sciences
Full Address (for proofs only):
1036 N. College Ave.
Claremont, CA 91711

Key words (list 10-15 in alphabetical order):
Educational evaluation, Program evaluation

Glossary:
No entries
Abstract

Program evaluation is the systematic study of curriculum, programs, and policies to address questions related to their merit, worth, value, and impact. In education, this process includes the examination of program processes and outcomes for the purpose of making both formative and summative judgments. We describe an evaluation framework that offers a clear description of the steps involved in designing and conducting program evaluations including: engaging stakeholders, describing the program, focusing the evaluation, gathering credible evidence, justifying conclusions and ensuring lessons learned.
Evaluation is the process of characterizing and appraising something of interest. In everyday life, we evaluate naturally. For example, we make evaluative judgments about the type of milk we buy in relationship to a set of standards – fat content, calories, taste, etc. This type of “every day” evaluation is distinct from formal evaluation, which involves the systematic study of some evaluand (the object of the evaluation) to address questions related to the merit, worth, value, and impact of the evaluand. While in education we evaluate personnel, products and programs, most of the theoretical, conceptual and practical evaluation work that has been done over the years is grounded in evaluation of programs. Thus, the field that is now known as evaluation has evolved primarily from program evaluation.

In education, program evaluation is mostly concerned with the study of curriculum, programs, and policies. These “programs” can be examined in relation to their process, outcomes and impact. The primary purpose of a process evaluation is to examine the development and implementation of the program. This information helps to explain what program components worked and why. Additionally, this information is useful for monitoring program implementation, and more generally for program improvement. A process evaluation focuses on what services were provided to whom and how, describes how the program was implemented, who was involved and what problems were experienced.

Gomby and Larson describe the evaluation of a complex school-linked service collaboration among a number of child-serving agencies and suggest several roles for a process evaluation. For example, a process evaluation might examine how the agencies interacted during implementation and document changes in relationships among social
and health service providers, educators, and client families. A process evaluation might also document system change as evidenced by new intake procedures, new forms, memoranda of understanding, or interagency linkage agreements. In some cases, changes in relationships and systems may be specifically planned goals of the program. In other cases, such changes may only be intermediate outcomes accomplished en route to the outcome of most interest to the program: changes in the students.

Outcome evaluation examines the program’s benefits to recipients; the direct effects of the program on participants. For example, an outcome evaluation would examine whether participants in a 3-month program designed to teach adolescents to practice safer sex can demonstrate the skills successfully. This type of evaluation is not unlike what happens when a teacher administers a test before and after a unit to make sure the students have learned the material. The scope of an outcome evaluation can extend beyond knowledge or attitudes, however, to examine the immediate behavioral effects of programs.

Impact evaluations look beyond the immediate results of the program to identify possible longer-term effects. An impact evaluation typically examines whether a program's immediate positive effects on behavior were sustained over time. Many school districts limit their evaluations to process and outcome studies. However, there are other agencies that are interested in and have the resources to examine whether program activities affect participants over time. For example, in the U.S. there has been a desire to understand the longer-term impact of early childhood education programs such as Head Start on students’ academic skills over time. Thus, academic outcomes of students who attended Head Start have been studied through high school and beyond.
A significant proportion of educational program evaluation takes place in traditional school settings (i.e., K-12 schools and higher education). In the U.S., federal, state and local departments of education support a great number of studies investigating the impact of educational programs so that policy and other decision makers can modify or enhance existing programs and initiate new education initiatives. Information generated from educational program evaluations can be used formatively, meaning that the information generated is used to improve programs on an ongoing basis. Summative evaluations may also be conducted, generating information that enables schools to determine, for example, which reading curriculum has the greatest impact on student literacy scores. Hence, summative evaluation conducted in the school setting is an “endpoint” evaluation of a particular program that has specific and expected outcomes.

Ralph Tyler conducted one of the first large-scale educational evaluations in the U.S.. Known as The Eight Year Study (1933-1941), this longitudinal study of progressive education through the four years of secondary school and the four years of college is still considered to be one of the most significant evaluations of the 20th century. Tyler’s work on this study has been described as one of the best available examples of how evaluators can work cooperatively with teachers to clarify instructional objectives and develop indicators of students’ continuous progress on a range of learning outcomes.

While other important studies of school-based programs were conducted during the 1940’s and 50’s, it was not until the Great Society legislation of the 1960’s that a broad and coordinated effort to evaluate federal programs was supported. This legislation introduced new programs into U.S. classrooms and appropriated funds specifically for evaluating their impact. The same stipulation was placed upon other large social
initiatives designed to support people and the communities in which they lived (e.g., Aid to Families and Dependent Children (AFDC)). This new demand for evaluation served as the foundation for the professionalization of evaluation, which continued to grow and mature during the 1970’s. Reductions in federal spending during the 1980’s translated into less funding for program evaluation and, as a result, a decline was experienced in the number of program evaluations conducted. This changed however in the 1990’s, during which the U.S. federal and state government spending for evaluation increased. This increase in available funds for program evaluation resulted in renewed energy for studying social and educational programs. This energy continues today.

**Defining program evaluation: What it is and how it is different from research and assessment**

The evaluation literature includes a diversity of opinions about the ultimate purpose of program evaluation. While these differences are subtle, they are meaningful and become clearer when reviewing how some leading academicians describe program evaluation in their texts. For example, Michael Patton defines evaluation as the systematic collection of information about the characteristics, activities and outcomes of programs to improve program effectiveness, make judgments about the program, or inform decisions about future programming. Peter Rossi, Mark Lipsey, and Howard Freeman highlight the importance of using information generated from evaluations for decision-making, but also emphasize the role of evaluation in promoting social betterment. Others focus on the use of evaluation to identify causal outcomes of programs, that is, the systematic assessment of the extent to which a program caused the observed results. Amid these differences one striking similarity emerges—the systematic
nature of evaluative inquiry. To this end, evaluators employ the same methodological tools used by researchers such as study designs, data collection, and analysis techniques. Yet, research and evaluation differ in important ways: the origin of study question(s) and the purpose for which study information is gathered.

With respect to the origin of study questions, evaluation questions are typically elicited from program stakeholders whereas research questions originate from a researcher or research team. That is, researchers hypothesize about phenomena, and then develop research studies to answer their question. Regarding study findings, evaluation generates information for program improvement and decision-making; that is, evaluation is decision oriented and is intended for use by individuals seeking the information. Evaluation studies, then, are designed to offer decision-makers, such as program directors and policy makers, systematic information about a program so that some actions can be taken, such as whether to continue a program. Alternatively, research is conclusion oriented. That is, the information generated from research studies is intended to enhance our general understanding of an area. The study of learning and cognition, for example, includes consideration of basic psychological theory, educational issues and application. By and large, researchers generate hypotheses about phenomena with the purpose of producing information that can be generalized across context, people, and time, whereas evaluators typically serve an identified group of stakeholders (e.g., program staff) with the goal of providing information that will be useful for making decisions about specific programs within a relatively defined timeframe.

Evaluation is also different from assessment. While the terms assessment and evaluation are often used interchangeably in our everyday vernacular, in the context of
education, assessment and evaluation are two distinct (although related) activities. Assessments are tests that are developed and administered to measure student performance. They are often used to measure specific student outcomes in the context of a broader educational program evaluation. An assessment system consists of some combination of norm-referenced, criterion-referenced, alternative, and classroom assessments reported to inform decisions about students, schools, districts, or states.

What distinguishes an assessment system from the more general program evaluation is the focus on testing; assessment systems use tests as the basis for decision making. In evaluation, information about program processes and outcomes is generated for decision making using any combination of quantitative and qualitative data collection techniques, which may or may not include testing.

**Models and Frameworks for Program Evaluation**

In 1967 Edward Suchman, a sociologist at the University of Pittsburgh, wrote a founding book on evaluation research citing Campbell and Stanley’s 1963 manuscript, *Experimental and Quasi-Experimental Designs for Research* as “the appropriate guide for developing evaluation designs.” This citation brought experimental and quasi-experimental designs to the center of evaluation. Given that most of the evaluations at the time were being conducted by university based social scientists, these designs were valued as “scientific” and thus, desirable. Rather quickly, however, many individuals conducting evaluations began to recognize that, for a variety of reasons, experimental designs were very difficult to implement in educational settings. The school context, and the districts in which they operated, were not necessarily conducive to the Campbell & Stanley designs, particularly the randomized experiment. Evaluators acknowledged that
in many contexts experiments are not achievable, ethical or desirable. Even the alternatives to the true experiment, quasi-experimental designs, that were developed to deal with the “messy” world of field research, were not always practical or popular. Alternative approaches for conducting evaluation were developed in response to these critiques and observations. Today a diverse set of ideas and beliefs exist about how to best conduct evaluation in the educational setting.

The vast majority of evaluation approaches (referred to as models, frameworks, or theories) offer principles, rationales, and organizational structures for the procedural choices made by evaluators thereby orienting them to the issues and problems with which they must deal. By and large, they are qualitative models, points-of-view, persuasions, and approaches to the process of evaluation, which are intended to guide practice rather than explain phenomena. The U.S. Centers for Disease Control and Prevention (CDC) developed a framework for program evaluation with input from hundreds of professionals and practitioners working in evaluation and research in universities as well as local communities. This framework describes the general steps involved in designing and conducting evaluations of social and educational programs. Although developed by a public health agency, the steps are generic and easily adapted for use in a variety of program contexts, particularly educational settings.

The framework consists of six-steps: 1. Engage stakeholders, 2. Describe the program, 3. Focus the evaluation, 4. Gather credible evidence, 5. Justify conclusions, and 6. Ensure use and share lessons learned (See Figure 1). During each of these steps, evaluators are encouraged to apply the program evaluation standards of utility, feasibility, propriety, and accuracy as described by the Joint Committee on Standards for
Educational Evaluation. These standards assist evaluators in making difficult decisions encountered in evaluation practice, therefore it is anticipated that use of these standards will result in high quality evaluations. The linear arrangement of these steps is intended to clarify how program evaluations are structured and to emphasize the importance of fully addressing each step prior to making final determinations in subsequent steps. However it is recognized that some of the steps residing at the end of the framework may be addressed earlier in the process (e.g., how the evaluative findings will be used is often considered when engaging stakeholders).

<INSERT FIGURE 1 ABOUT HERE>

Engaging stakeholders

Stakeholders are people who have a vested interest in the evaluation findings. However, the level of interest stakeholders have in the program or its evaluation can differ. For this reason, stakeholders are sometimes categorized based on their relationship to the evaluation: primary stakeholders are in a position to use findings from an evaluation to alter a program’s course (e.g., program designers, program staff); secondary stakeholders are likely to be affected by programmatic changes (e.g., teachers, parents, students), and tertiary stakeholders may be interested in the findings from the evaluation but are not directly impacted by its results (e.g., superintendents of other school districts interested in adopting the program).

Two key considerations associated with stakeholder engagement include when to involve stakeholders and which stakeholders to involve. Program evaluators vary in the breadth with which they engage stakeholders—some evaluators focus solely on including primary stakeholders whereas others make extensive efforts to engage secondary and
tertiary stakeholders. When making determinations about which stakeholders to involve, it is helpful to consider how non-traditional partners may contribute to the evaluation process. For example, it may be tempting to limit stakeholder involvement to individuals who directly influence the program (e.g., program managers), yet their experiences may differ greatly from those who receive program services. It may also seem counterintuitive to involve individuals who are not supportive of the program. However, engaging these non-traditional evaluation stakeholders may result in a richer program description, identification of important evaluation questions that would have been overlooked, and lend greater credibility to evaluation findings to a broader audience. Despite the potential value that each stakeholder perspective can bring to the table, evaluators need to carefully balance the value of adding more stakeholders with the need to make progress in the evaluation.

Program evaluators also vary with respect to the depth with which they engage stakeholders. Some evaluators involve stakeholders in every step of an evaluation, including data collection, analysis, and interpretation whereas others limit the scope of stakeholder involvement to the early stages of an evaluation (e.g., defining the program and determining evaluation questions). This variation in practice is a result of evaluator characteristics (e.g., the evaluator’s beliefs about the ultimate purpose of evaluation, the evaluator’s training) and logistical considerations associated with the evaluation (e.g., deadlines, funding levels, access to stakeholders).

Describing the program

Studies of evaluation practice show that it is common for program evaluators to engage stakeholders in developing a description of the program. A common technique
used to do this is logic modeling, in which a pictorial description is developed with stakeholders to clearly articulate how program activities will lead to fulfillment of programmatic goals. A general depiction of a logic model is provided in Figure 2.

When developing a logic model, evaluators may review existing program documents, conduct observations, and meet multiple times with stakeholders to obtain information from various perspectives about what constitutes a successful program. Logic models vary in their composition and structure, however common components include inputs, activities, outputs, as well as short-term, intermediate, and long-term outcomes. Logic models sometimes also include a description of programmatic assumptions and denote the external forces that have the potential to impede or enhance progress towards reaching the stated program outcomes.

<INSERT FIGURE 2 ABOUT HERE>

While developing the logic model, it may be discovered that the theory underlying the program is underdeveloped, unclear, or is unlikely to produce the anticipated results based upon findings from previous research studies. In these instances, the program may be modified prior to embarking upon an evaluation. In the event that the evaluation continues, as is typically the case, the presumed causal pathways expressed in the logic model can be helpful in generating appropriate and useful evaluation questions, selecting an evaluation design, identifying appropriate time periods to collect data on the variables in the model, and ultimately in communicating findings. Logic modeling is only one approach to describing a program, related techniques include but are not limited to program theories, action models, and impact models; all of which are discussed as part of an approach to evaluation known as “theory driven evaluation”
and described in detail by evaluation scholars such as Peter Rossi, Huey-Tsyh Chen, and Stewart Donaldson.

In addition to developing a logic model, or a similar programmatic representation, evaluators may also work with stakeholders to elucidate and document additional factors that are important to fully describing the program. Specifically, it is important to place the description of the program itself within a broader context. Clearly articulating the need(s) addressed by the program, the context in which it resides (e.g., budgetary considerations, political considerations), the length of time the program has been in operation (i.e., whether the program is in a planning or implementation phase), and the degree of interconnectedness that exists between this program and related endeavors can help the evaluator and stakeholders in designing, implementing, and interpreting the findings and implications of the program evaluation.

Focusing the evaluation

When focusing the evaluation, evaluators work with stakeholders to clarify the intended purpose of the evaluation, generate a list of clearly articulated evaluation questions that align with the intended purpose, prioritize the questions on this list, and select an appropriate evaluation design. This process requires a thorough understanding of why the evaluation is being commissioned and how stakeholders who are in a position to use evaluation findings actually intend to use the findings. During this step, the evaluator works with stakeholders to carefully balance considerations based upon the program evaluation standards of utility, feasibility, propriety, and accuracy.

As previously mentioned, the logic model developed with stakeholders can be used to generate a list of potential evaluation questions. As part of this process it is likely
that a number of evaluation questions will be raised however practical considerations, such as the length of time the program has been in operation and the resources available to conduct the evaluation (i.e., staff and money), make it unlikely (and in some cases inappropriate) to address all of the questions proposed. After a limited number of meaningful and useful evaluation questions have been prioritized with the stakeholders, an evaluation design that is capable of addressing these questions is selected. Evaluation designs include those that are experimental (e.g., randomized controlled trials), quasi-experimental (e.g., pretest-posttest control group), or non-experimental (e.g., ethnography) in nature.

Gathering credible evidence

During this stage of the program evaluation consideration is given to the types of data that will be collected as part of the study as well as the specific data collection methodologies that will be employed. In the previous step, the evaluator worked with the stakeholders to articulate the purpose of the evaluation and to generate and select a set of clearly worded evaluation questions. The data collected should contribute directly to answering the evaluation questions selected and should produce credible evidence to answer them.

Data collection strategies used by the evaluator may include the collection of new data (i.e., primary data) through such methods as surveys, interviews, observations, or focus groups. Evaluators may also choose to use data that is already available (i.e., secondary data), which might include archival information from existing electronic or paper records, or analyzing data that is used to monitor, for example, student progress (e.g., student assessment data). The credibility of data gathered through these methods
can be affected by the procedures used to train individuals in data collection, the validity of the instruments used to gather data, the reliability of the measures used, and the quality control procedures used to monitor data integrity and accuracy. Additionally, it is important for evaluators to adhere to ethical standards associated with the protection of data, and to maintain confidentiality and anonymity in instances where these protections have been assured.

The credibility of evaluation evidence is, at least in part, in the “eye of the beholder.” To develop an evaluation that is regarded as credible to the stakeholders, many evaluators will work with stakeholders to gain a clear understanding as to what constitutes credible evidence from their perspectives. Stakeholders may have a preference for the type of design selected in the previous step (e.g., view randomized controlled trials as the “gold standard”) as well as the type of data that is collected for the evaluation (e.g., qualitative, quantitative, or both). Engaging stakeholders in discussions to better understand the expectations they hold about what types of evaluative data will be collected is an important step in an evaluation, and can help to increase the likelihood that evaluative findings will be used.

**Justifying conclusions**

After data have been collected the next step is to analyze, synthesize, and interpret the findings from the evaluation. To prepare for this step, evaluators will often work with stakeholders during the design phase to determine what level of achievement constitutes programmatic success. These shared value statements form the standards that can be used to judge the evidence produced through the evaluation. When interpreting the findings, the evaluator and/or stakeholders may embark on a process in which they critically
examine the proposed reasons for why the specific evaluative findings emerged and attempt to find alternative rationales (e.g., plausible rival alternative hypotheses). Additionally, they may choose to see how their results “measure up” to similar indicators of program performance—such as their performance on the same indicator in a previous time period, or an aggregate measure of performance at a state or national level.

Irrespective of the method of interpretation used, it is important to consider what the practical implications of the findings are for the program and how the findings can be used to make changes or inform decisions about the future of the program. Any next steps or recommendations made during discussions with stakeholders should be firmly grounded in the analysis and synthesis performed on the credible evidence gathered in the previous step. Moreover, findings from a program evaluation should be applied in other settings with caution. Generalizability of the findings should be done in accordance with the limits imposed by the evaluation design and statistical sampling techniques employed (if any).

Ensuring use and sharing lessons learned

Although ensuring use and sharing lessons learned reside within the final step of the framework, considerations about how to facilitate the use of evaluative findings should begin early in the process of developing the evaluation and continue throughout the evaluation process. For example, when focusing the evaluation, it should become clear as to how the stakeholders ultimately anticipate using findings from the evaluation as well as their information needs. Thus, the evaluator should work early in the evaluation process to determine the types of decisions that might result from the evaluation. For example, will the information from the study be used to determine
whether the program will be continued or if it may be implemented in other geographic regions and if so what changes might need to be made to the program so that implementation elsewhere is successful. This interaction aids the evaluator in tailoring the evaluation in a manner that increases the likelihood that findings will be used when the evaluation is complete.

Other steps can be taken throughout the evaluation process to prepare stakeholders for using the evaluation results. For example, interim findings or information about the progress that is being made in the evaluation may be shared with the appropriate stakeholder group(s). These interactions can help to prevent surprising the stakeholders with unanticipated findings, which may lead to stakeholders rejecting the final results. Additionally, the evaluator can engage stakeholders in “mock” scenarios while the evaluation is underway. During these scenarios, the evaluator presents fallacious, but realistic, evaluative findings to the stakeholders and works with them to think through how they would interpret these findings and take action.

Another technique used to facilitate the use of evaluation findings is the development of a communication and reporting plan that is tailored to stakeholder needs. A communication and reporting plan clearly describes when, how, and with whom communications about the evaluation will be shared and identifies the specific methods that will be used to communicate with stakeholders. Potential communication methods include but are not limited to hard copy reports, presentations, emails, working sessions, and newsletters. As is demonstrated through these examples, communications range in their level of formality (i.e., informal emails to formal presentations) as well as in their level of interaction (i.e., highly interactive working sessions to static newsletters).
Summary/Conclusion

The evaluation of social and educational programs has an important role in understanding which programs are effective and why. A range of approaches have been developed to help guide evaluators in the conduct of program studies, which are intended to generate reliable, valid, and relevant information for decision-makers to use when determining which programs to revise, expand, or terminate. To this end, it is critical that evaluation information is timely, meaningful, credible, and accessible to all that have a stake in both the program and its evaluation.

Bibliography


**Further Reading**


Websites of use to reader

American Evaluation Association:

www.eval.org

Center for Disease Control and Prevention’s Evaluation Working Group:

http://www.cdc.gov/eval/index.htm

Western Michigan University: The Evaluation Center

http://www.wmich.edu/evalctr/
Suggested cross references:

Chapter 5: Formative and summative evaluation
Chapter 9: The history of evaluation
Chapter 11: Moral and ethical issues in evaluation
Chapter 12: Cultural/multicultural considerations in evaluation
Chapter 16: Evaluation use/utilization
Chapter 18: Evaluation methodology
Chapter 23: The role of stakeholders in evaluation
Chapter 24: Setting evaluation standards
Chapter 26: Evaluation reporting and communicating

Biographies:

Christina A. Christie:

Christina A. Christie is Associate Professor in the School of Behavioral and Organizational Sciences at Claremont Graduate University. Christie co-founded the Southern California Evaluation Association, a local affiliate of the American Evaluation Association, and is the former Chair of the Theories of Evaluation Division and current co-Chair of the Research on Evaluation Division of the American Evaluation Association. In 2004, Dr. Christie received the American Evaluation Association’s Marcia Guttentag Early Career Achievement Award. Christie is a section editor of the American Journal of Evaluation and on the editorial board of New Directions in Evaluation.

Leslie A. Fierro:

Leslie Fierro earned her Bachelors of Arts degree in Biology from Pitzer College and her Masters of Public Health degree in Epidemiology and Biostatistics from Loma Linda University. She holds a Certificate of Advanced Study in Evaluation from Claremont Graduate University. Currently, she is continuing her academic endeavors in evaluation at Claremont Graduate University where she is pursuing a PhD in Evaluation and Applied Research Methods. Previously she was employed with the Centers for Disease Control and Prevention’s National Center for Environmental Health, where she became interested and involved in program evaluation.
Figure 1 – CDC Evaluation Framework