EVALUATION BRIEF

Selecting an Evaluation Approach

September 2009

Introduction

Program evaluation is an essential component of nearly all child welfare discretionary grants funded by the U.S. Department of Health and Human Services. Evaluation involves the systematic collection, analysis, and use of information to answer basic questions regarding the overall effectiveness of a program or about specific services or activities implemented through the program. The term *systematic* denotes the use of structured and consistent methods for collecting and analyzing information. Any systematic evaluation effort requires the selection of an overarching research design to guide data collection activities and to ensure that they are implemented in a reliable and coherent manner. This brief discusses programmatic and contextual factors to consider when choosing an evaluation approach, reviews the most common types of evaluation designs, and offers a critique of common myths and misconceptions regarding various evaluation designs and research methods.

What is an Evaluation Design?

An evaluation design[1] refers to the overarching methodological framework that guides an evaluation effort; in other words, it is the conceptual lens through which the evaluation is viewed and implemented. The research design "provides the glue that holds the research project together. A design is used to structure the research, to show how all of the major parts of the research project … work together to … address the central research questions."[2]

Factors to Consider in Selecting a Design

Evaluations, like any public policy or human service endeavor, do not operate in a vacuum; as such, the selection of an evaluation design is not simply a matter of identifying the most rigorous and "scientific" research methodology. Rather, an evaluator must consider the

[1] Throughout this paper, the terms "research design", "evaluation design", and "evaluation approach" are used interchangeably unless otherwise noted.
[2] Trochim, W. M., 2006. Research methods knowledge database. Retrieved June 26, 2009, from http://www.socialresearchmethods.net/kb/design.php.

available resources and a range of organizational and other contextual variables that affect the implementation of an evaluation in real-life settings. With these pragmatic considerations in mind, the overarching goal is to select and implement a research design that results in the highest-quality and most credible findings possible given the objectives, design, and core features of the program in question. This synthesis describes the recommended steps in addressing these variables, which in turn influence the eventual choice of an evaluation design.

Step 1: Define Your Evaluation Objectives

The first step in selecting an optimal design involves thinking through the ultimate objectives of your planned evaluation. Why are you undertaking the evaluation and what goals do you hope to achieve as a result of the effort? For example, do you wish to:

- Choose between different programs or service alternatives?
- Improve or modify a new or existing program or service?
- Inform your profession or academic field about a new effective practice or service model?
- Make budgetary or funding decisions?

Clarifying the purpose(s) for which the evaluation will be used can assist you in delineating the required scope and level of effort required to implement your evaluation, as well as the degree of rigor and integrity that will be expected from your evaluation findings.

Step 2: Define the Parameters of Your Evaluation

Once you have a clear idea of what you wish to accomplish through your evaluation, it is essential to review a range of contextual factors that will further delimit the timing and scope of evaluation activities. Major questions you should ask yourself that will assist in defining the parameters of the evaluation include:

- When will you need the evaluation findings?
  - Do you have time to collect data *prospectively* versus *retrospectively*?
- What kinds of data are available?
  - Are *case-specific* or only *aggregate* data available?
  - Are *baseline* data available regarding services and key outcomes for your target population?
  - Will you need to design your own data collection instruments to address your primary research questions or is appropriate information available from existing data sources (e.g., child welfare information management systems)?
  - Can you track data on services that any proposed control or comparison group receives?
- What resources are available to conduct the evaluation (e.g., money, staff, facilities)?
  - Do you have internal staff that can conduct the evaluation or will you need to rely in whole or in part on an external third-party evaluator?

The italicized terms noted above are of particular importance in assessing the types of information that will be available for your evaluation effort:

- In a *prospective* evaluation, implementation of a new program or service has not yet occurred and achievement of the desired outcomes is expected at some point in the future. Consequently, new data must be collected to determine when and whether the outcomes of interest actually occur. Because hypothesized outcomes are predicted to happen at some indeterminate point in the future, data collection may need to continue for a lengthy period. In contrast, *retrospective* studies usually involve programs, policies, or services that have already been implemented; therefore, the outcomes of interest may have already occurred and the evaluator must look back in time by using existing data to determine whether the hypothesized outcomes were actually achieved. Since retrospective studies often involve the analysis of existing data, they usually require less time to plan and implement than prospective studies.

- Both prospective and retrospective studies may involve the analysis of *baseline* data, which includes initial information about the population of interest that is collected before implementation of the new program begins. Baseline data serve as a point of reference or benchmark for comparison with data collected after program implementation. In prospective studies, baseline data must be collected before program implementation begins, which adds additional time and cost to the evaluation. Baseline data collection in retrospective studies usually involves little additional time or cost; however, the baseline data themselves are only useful to the extent they are available, complete, and correspond to the outcomes you wish to track.

- *Case-level* data refer to demographic, service, and outcome information that is collected, tracked, and reported on each individual case. In contrast, *aggregate* data are collected, tracked, and reported for the entire group or population of interest. Case-level data are generally preferable because they allow an evaluator to correlate specific demographic or case characteristics with observed outcomes. For example, with case-level data an evaluator can assess whether the impacts of a family reunification program differ by the mental health or substance abuse histories of enrolled parents. Aggregate data, on the other hand, only allow for the examination of general trends or patterns across the entire population of interest; more sophisticated analyses of subsets of this population are usually not possible.

Step 3: Formulate Key Outputs and Outcomes

Having identified the major objectives and parameters that will define your evaluation effort, you can formulate specific measures that correspond to the two major components of an evaluation effort: the *process* evaluation and the *outcome* evaluation. The process component of an evaluation seeks to determine whether the program or service in question was actually implemented as intended. Specifically, it describes *who* received services as well as *how much* and *what types* of services were provided.[3] In contrast, the outcome

[3] For more information on process evaluation, see James Bell Associates (2008, August). *Conducting a process evaluation*. Arlington, VA: Author.

component of an evaluation asks whether changes in *knowledge*, *attitudes*, *skills*, *behaviors*, or *status*[4] were achieved as a result of the program. In other words, did the program realize the desired changes in its target population?[5] All comprehensive evaluations should include both process and outcome components. Examples of questions that a process and outcome evaluation might ask regarding a hypothetical child welfare program are provided in Table 1 below.

Table 1: Examples of Process and Outcome Evaluation Questions

Process Questions	Outcome Questions
▪ How many services were provided? ▪ How many people were served? ▪ What are the characteristics of participants (e.g., race, age, gender, in foster care)? ▪ How often were services provided (frequency)? ▪ How long did families participate in services (intensity)? ▪ How closely did actual services correspond to the original service model (fidelity)? ▪ How satisfied were participants with services?	▪ Did caregivers' parenting knowledge and coping skills increase? ▪ Did more children exit foster care to permanency? ▪ Did children spend less time in foster care? ▪ Were children less likely to experience repeat maltreatment? ▪ Did children's school performance improve?

By articulating your core process and outcome measures, you will further narrow and refine the scope of your evaluation efforts, which in turn will assist you in identifying the most effective and appropriate design.

Step 4: Review and Select a Design

Once the purpose, parameters, and core process and outcome measures for your evaluation have been identified, you can proceed with the selection of a research design. Regardless of its particular strengths and weaknesses, every good evaluation design involves the use of a comparison group or other point of reference that allows you to attribute observed changes to your new intervention and not to other unrelated factors. Although an exhaustive review of all major designs is beyond the scope of this paper, an important first step involves identifying the major features and approach of each design option; the initial circumstances under which a particular research design is a good alternative; and the conditions, resources, and procedures that must be in place to implement a design effectively. Using this assessment framework, this section provides an overview of the most common design alternatives.[6]

[4] "Status" refers to the long-term condition or circumstances of a person. For example, the permanency status of a child in a family reunification program is either reunited with her family of origin, still in foster care, or in another permanent living arrangement (e.g., adoption, guardianship).
[5] For more information on outcome evaluation, see James Bell Associates (2008, December). *Conducting an outcome evaluation*. Arlington, VA: Author.
[6] See Table 3 at the end of this brief for a summary of the most common design alternatives.

Experimental (Random Assignment) Designs

In experimental research designs, cases are randomly assigned to an "experimental" group (eligible for the new program or service) or a "control" group (ineligible for the new program/service). Sometimes referred to as the "gold standard" for social science research and evaluation, experimental designs have compelling advantages over other evaluation approaches:

- It is easier to attribute observed changes to the new program you are evaluating. Because the randomization process creates two groups that are essentially identical — except that the experimental group receives the new service in question whereas the control group does not — there are fewer extraneous internal or external factors that might explain observed participant outcomes.

- Although more difficult to set up on the front end (e.g., developing and implementing a random assignment protocol), data analysis is much easer on the back end because you have two distinct groups with little ambiguity regarding who received the new treatment or service and who did not. Since there is less "noise" in the data (caused, for example, by inadvertently exposing a person assigned to the control group to the experimental treatment), it is much easier to detect meaningful differences between the experimental and control group, and consequently, to attribute observed changes in children and families to the new program or service in question.

An experimental design is a preferred alternative when you have a discrete and clearly defined treatment or service and when there are more people potentially eligible for the new service than there are resources to provide the service to all eligible recipients. Experimental designs are also a good choice when the scientific credibility of your evaluation is paramount, which may be important for convincing some funders of the value and effectiveness of your new program.

Effective implementation of an experimental design requires buy-in from program management and staff to random assignment and an evaluation team with a high degree of technical expertise in quantitative research methods. In addition, a large sample size (*N*) is often required to detect meaningful differences in observed outcomes, and stringent procedures must be in place to prevent "spillover" or design contamination (a situation in which the control group is either inadvertently or deliberately exposed to the new program).

Although experimental research designs generally do the best job of isolating the actual impact of your new program on key child welfare outcomes as opposed to other factors, the nature and design of certain programs may preclude the use of random assignment. For example, large-scale neighborhood or citywide initiatives cannot always be evaluated using experimental designs because an entire community is by definition the target for service delivery and program resources are diffused throughout the entire population. In addition, a variety of practical and contextual factors (e.g., cost, lack of buy-in among project staff to random assignment) may make an experimental design infeasible. In these situations, several other design alternatives may be appropriate, each of which is described in more detail below.

Waitlist/Overflow Design

When an experimental design is not feasible, a waitlist/case overflow design can offer a good alternative. When caseloads for the new program have reached capacity and workers cannot accept new cases, people who are otherwise eligible for the new program are placed on a waiting list. Cases on the waiting list serve as the design's comparison group, with outcomes for waitlisted persons compared with outcomes for cases that are enrolled in the new program (the experimental group). As with experimental designs, waitlist/overflow designs work best with a discrete and clearly defined intervention and when there are more people potentially eligible for the new service than there are resources to provide the service to all eligible cases. Specific circumstances in which this design may be especially suitable include when there is limited support among program management and staff for an experimental design and a lack of technical expertise to implement random assignment.

To prevent design contamination, waitlist designs are most appropriate with projects of limited duration so that cases on the waiting list are not enrolled in the new program before the evaluation is complete. In addition, strict control over the assignment process is essential so that enrollment into the new program occurs on a "first-come first-served" basis rather than by relying on workers' subjective assessment of the needs of specific people or families. In this way, the characteristics and severity of cases in the comparison group remain relatively similar to those in the experimental group so that the waiting list continues to serve as an acceptable reference point for comparing child and family outcomes. Even with these safeguards in place, waitlist designs have methodological limitations that evaluators should be aware of, including the possibility of selection bias (for example, if control group clients need to be removed from the waitlist and provided services because of a significant deterioration in their well-being). In addition, because the waitlisted control group usually receives the experimental treatment at the end of the study, the design does not allow for long-term follow-up assessments, thus precluding the possibility of examining the differential effects of treatment over extended periods.

Matched Case Designs

Use of an experimental design is sometimes not possible because specific circumstances require an organization to offer a new program to as many people as possible, for example, when the target population for the service is small or when a contract with a service provider stipulates that a certain number of people must be served. In these situations, a matched case design serves as a good choice. With matched case designs, each case that is offered the new program (the experimental group) is individually matched with a comparison case based on selected matching variables (e.g., presenting problems, demographic characteristics). For example, matching variables for an out-of-home placement prevention program might include maltreatment type, maltreatment risk severity score, and age of child (among other criteria). Matched case designs include those that use propensity score matching (PSM), a statistical technique in which cases are matched using a composite score generated by an algorithm that minimizes variance across any one matching variable.

Matched case designs work well when your new program is (1) targeted at a discrete population of limited size, and (2) when a group of matching cases exists that has very similar presenting problems and case characteristics as the experimental group and that has not been exposed to the new program or to a similar intervention. To facilitate the matching process,

comprehensive and detailed information on the presenting problems and demographic characteristics for both experimental and matching cases must be available in some readily accessible format, ideally in a child welfare information system or similar database. Matched case designs that employ PSM generally require a somewhat larger sample size (N >= 200) to detect meaningful differences between the experimental and matched comparison group.[7]

Comparison Site Designs

The research designs described above may not be feasible when implementing a community- or system-wide reform rather than a discrete service targeted at a limited pool of cases. For example, in the case of a neighborhood-based child maltreatment prevention program that relies on community outreach and education (e.g., public service announcements, community events), services are diffuse and accessible to large numbers of people throughout the entire neighborhood. For such systemic interventions, organizations may choose to implement a comparison site design in which a community is identified that has characteristics similar to those of the target community in which the new program is implemented; differences in outcomes between the experimental and comparison community are then tracked and compared over time. Comparison designs are most effective when baseline data exist for both the experimental and comparison communities to allow for the measurement of rates of change in key outcomes.

Comparison designs are often a popular design choice because they require less effort to set up and are more widely accepted by program managers and front-line staff who may be unfamiliar and uncomfortable with experimental research designs. However, comparison designs have significant limitations that should be considered carefully before they are selected over other evaluation alternatives:

- It is difficult, and sometimes virtually impossible, to identify a community or political jurisdiction (e.g., city or county) that has the same characteristics (e.g., demographic makeup, socio-economic issues, child welfare case mix) as the target community. Consequently, any observed differences in outcomes between the target and comparison community may be due to pre-existing differences in these characteristics rather than a result of the new program of interest.

- It is unlikely that the comparison community will have *no* programs or reform efforts similar to those implemented in the target community. This leads to another type of design contamination, in which people in the comparison community are exposed to services or activities that are similar to those implemented in the target community. As a result, people in the comparison community may exhibit some of the same changes as those in the target community, rendering it even more difficult to isolate the unique effects of the new program you are evaluating.

- Even when a suitable comparison site can be identified, it is not always certain that equivalent process and outcome data for the comparison community exist or that they

[7] For more detailed information on propensity score matching, see Rosenbaum, R., & Rubin, D. (1983). The central role of the propensity score in observational studies for casual effects. *Biometrika*, *70*(1), 41-55; and Pearl, J. (2009). Understanding propensity scores. In *Causality: Models, Reasoning, and Inference* (2nd ed.). Cambridge, UK: Cambridge University Press.

are available in a form that allows for meaningful comparisons. In addition, gaining access to the requisite data may depend on the cooperation of authorities in the jurisdiction that will serve as a comparison site; these authorities are often focused on other priorities and may have little incentive to assist with an evaluation in which they have little stake and from which they will derive limited benefit.

Time Series

Sometimes no suitable community or group of people can be identified to serve as a reference point for measuring differences in child welfare outcomes. In these circumstances, time itself can serve as the basis for comparison, with changes tracked longitudinally for the same program. With such "time series" designs, baseline data are compared at regular intervals (e.g., every six months) with information collected during and after program implementation. Longitudinal data are often tracked in "entry cohorts," for example, all children who enter foster care in a given year.

In addition to being a viable alternative when no suitable comparison group can be identified, time series designs work best with programs that are implemented over an extended period of time (often many years) and for which only historical baseline data are available.[8] Because no external comparison group exists to account for factors that might explain changes in observed outcomes (other than the new program itself), the successful implementation of a time series design also depends on extensive knowledge of programs, policies, or changes in law that may occur in the future and that could affect key outcomes of interest.

Pre-Post Test

Some programs (e.g., classroom-based training) focus primarily on knowledge acquisition or changes in attitudes and short-term behaviors. These types of programs can often be evaluated using another longitudinal design known as a "pre-post test" design. With pre-post test designs, data on program participants are collected once at program entry (baseline data collection) and then one or more times thereafter, usually at program exit followed by another designated interval (e.g., six months after program completion). Data collection does not usually extend beyond this point.

As with time series designs, pre-post test designs lack an external comparison group to serve as an independent reference point, and they are not well suited to measuring long-term changes in people's behavior or status. However, they can offer a suitable alternative when evaluating training programs (and other programs focused on short-term change) and when time, resources, and technical expertise to implement a more sophisticated research design are limited. In addition, pre-post test designs are most effective when used to evaluate stable programs with minimal participant attrition (because a low post-test response rate will not generate enough data to allow for valid comparisons with pre-test data) and when they employ a standardized assessment instrument to further enhance the validity and reliability of evaluation findings. In addition, accurate participant contact information is helpful to optimize response rates if your pre-post test involves follow-up data collection with people who have completed the new program.

[8] See the discussion earlier in this brief about retrospective data.

Case Studies

A comprehensive evaluation seeks to identify not only what outcomes were achieved as a result of a new program but also *why* and *how* these outcomes occurred. Moreover, in the case of new programs or practice models there is often very little information regarding how these programs or models can be implemented most effectively or how they are experienced by program participants. In these situations, case studies can supplement an existing evaluation or serve as the basis for an independent research project. Rather than measuring results, case studies focus on understanding the *experiences* of people in a program and the *meaning* the program has for them. Moreover, unlike other research designs that often study a broad range of program activities and outcomes, the object of interest in a case study is limited to a single unit or bounded system (e.g., a specific person or event). Although quantitative methods can be employed, data collection in case studies most frequently involves qualitative research methods. Regardless of the methods employed, all case studies seek to gain a rich and in-depth understanding of the experiences of one or more program participants and of the social, cultural, and organizational context in which those experiences unfold.[9]

Due to the substantial investment of time and resources required to conduct *in situ* field work, case studies should not be regarded as a simple add-on to an existing evaluation. On the contrary, all good case studies require the services of an evaluator who is especially skilled in qualitative research methods such as conducting interviews, moderating focus groups, and coding and interpreting textual data.

Step 5: Review and Select Research Method(s)

A research method refers to a specific tool or technique that is used to collect data on selected process and outcome measures within the parameters established by the evaluation design. Although a chosen evaluation design often involves both qualitative and quantitative research methods, some of the same conditions and constraints that affect the choice of a research design also influence the selection of appropriate and practical research methods. Table 2 highlights some of the most common types of quantitative and qualitative methods.

Table 2: Examples of Common Research Methods

Quantitative	Qualitative
▪ Standardized assessment instruments and tests ▪ Surveys/questionnaires ▪ Analysis of existing administrative/IMS data ▪ Case record review (e.g., data on program attendance and service receipt) ▪ Structured observation (e.g., using numeric rating scales)	▪ Open-ended and semi-structured interviews ▪ Focus groups ▪ Document review (e.g., workers' case notes) ▪ Observation (e.g., taking detailed field notes or making journal entries)

[9] For more detailed information on case study research methods, see Marriam, S. B. (1998). *Qualitative Research and Case Study Applications in Education*. San Francisco: Jossey-Bass Publishers.

As you consider your alternatives, it is important to be cognizant of common misconceptions regarding certain evaluation designs and research methods. Some of the most prominent evaluation "myths" include the beliefs that experimental research designs are unfair and unethical because they deny people services, and that qualitative research methods are less rigorous and "scientific" than quantitative methods. Although there are many valid programmatic, methodological, and pragmatic reasons for selecting or rejecting a particular evaluation design or research method, these myths should not be among them. For a more detailed examination of common evaluation myths, see JBA's separate brief on this topic entitled *Common Evaluation Myths and Misconceptions* (James Bell Associates, 2009).

For more information about assessing and selecting an evaluation design, please contact a JBA team member at:

James Bell Associates
1001 19th Street, North, Suite 1500
Arlington, Virginia 22209
703-528-3230 or 800-546-3230
www.jbassoc.com

Table 3: Summary of Evaluation Design Alternatives

Design Type	Description	This is a Good Alternative If:	What you Need to Implement this Design Effectively
Experimental (Random Assignment)	Cases are randomly assigned to an experimental group (eligible for the new program or service) or a control group (ineligible for the new program/service)	▪ You have a discrete/clearly defined treatment or service ▪ There are more people eligible for the service than resources to provide the service ▪ Scientific credibility is paramount	▪ Buy-in from program management and staff ▪ High degree of technical expertise ▪ Large sample size (N) ▪ Controls and procedures to prevent "spillover" or design contamination
Waitlist/ Overflow	Cases are placed in a comparison group when caseloads for the new experimental program are full. Outcomes for cases in the waitlisted comparison group are compared with outcomes for cases that receive the experimental program	▪ You have a discrete/clearly defined treatment or service ▪ There are more people eligible for the service than resources to provide the service ▪ Support among management and staff for an experimental design is limited	▪ Treatment/service of limited duration ▪ You can prevent waitlisted families from receiving the new program/service until the evaluation is complete ▪ Strict control over the assignment process (assignment should be based on a "first come first served" rule rather than on workers' assessment of need)
Matched Case	Each experimental group case is individually matched with a comparison case based on selected matching variables (e.g., presenting problems, demographic characteristics)	▪ Your program/service is targeted at a limited and well-defined population ▪ You must provide the service to as many people as possible (e.g., because of small sample sizes, requirements of a service provider contract) ▪ You can identify a group of matching cases with similar presenting problems and case characteristics ▪ Support among management and staff for an experimental design is limited	▪ Comprehensive and detailed data on presenting problems, case characteristics, and demographics of both experimental and matching cases

Design Type	Description	This is a Good Alternative If:	What you Need to Implement this Design Effectively
Propensity Score Matching (PSM)	Type of matched case design. Cases are matched based on a composite "propensity score" to minimize differences across any one matching variable	▪ See matched case design above	▪ Large sample size (N>200) ▪ Comprehensive and detailed data on presenting problems, case characteristics, and demographics of both experimental and matching cases
Comparison Site	A community (e.g., neighborhood, city, county) is identified with characteristics similar to the target community in which the experimental program is implemented; differences in outcomes between the experimental and comparison community are tracked and compared over time	▪ You are implementing system- or community-wide reform rather than a discrete program, treatment, or service ▪ A community exists that has similar characteristics *and* no similar services or reform efforts in place	▪ Cooperation and assistance from authorities in other jurisdictions with data sharing and collection ▪ Ability to collect baseline data from both the experimental and comparison communities to measures rates of change in key outcomes
Time Series	Baseline data are compared at regular intervals with data collected during and after program implementation. Data are often tracked in "cohorts" (e.g., all children who enter foster care in a given year)	▪ Implementation of the program or service will occur over an extended period of time ▪ No suitable comparison group or site can be identified ▪ Only historical baseline data are available	▪ Knowledge of future initiatives, changes in laws or policies, etc. that could affect observed outcomes ▪ Adequate time for data collection
Pre-Post	Data are collected once at program entry and then one or two more times thereafter (usually at program exit or at some designated interval)	▪ You are implementing a training program or other project focused on knowledge acquisition or changes in attitudes and short-term behaviors ▪ Time, resources, and expertise to implement a more sophisticated design are limited	▪ Stable program with minimal participant attrition ▪ A standardized assessment instrument is recommended (to maximize the validity of results) ▪ Accurate participant contact information (if needed for follow-up data collection)
Case Study	The object of study is a single unit or bounded system (e.g., a person or event). Focuses on understanding the *experiences* of people in a program and the *meaning* it has for them. Relies most frequently on qualitative research methods	▪ You want to complement an existing research design to learn more about *why* and *how* certain outcomes occurred	▪ Substantial time and resources to conduct fieldwork ▪ Evaluator skilled in qualitative research methods (conducting interviews, moderating focus groups, coding and interpreting textual data)