

Chapter 5. Design Your Evaluation

What's Inside?



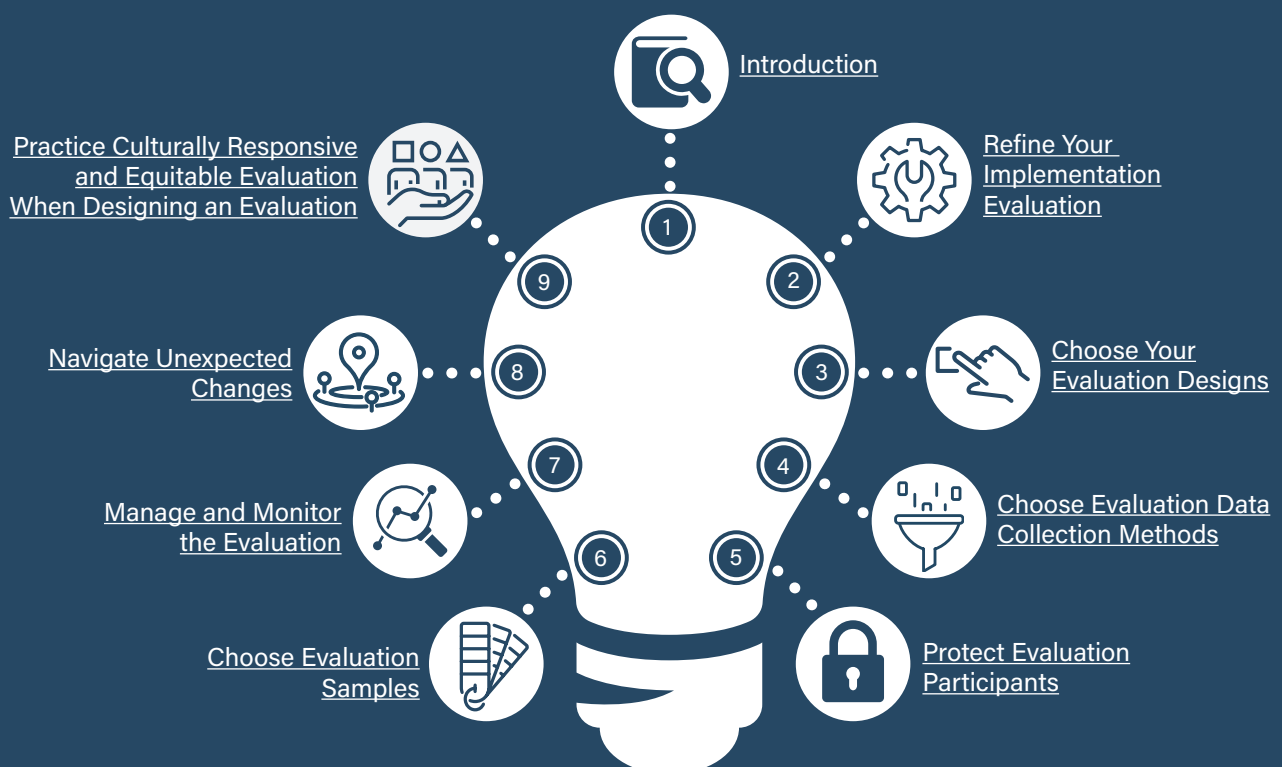
What this chapter contains

- A brief introduction to implementation and outcome evaluation designs and common data collection methods
- A discussion of evaluation management, including ways to protect evaluation participants

Who can use this chapter

- Evaluation team members preparing to develop an evaluation plan

Click the links below to view the relevant section



Introduction

Now that you have developed your evaluation questions, you need to determine how you will answer each one. That means some big decisions to develop an approach for each question:

- Determine the appropriate evaluation designs.
- Select the appropriate data collection methodologies.
- Establish each data collection's source of data.
- Identify the appropriate measures for each evaluation question's concepts.

This chapter covers the first two steps, and [chapter 6](#) addresses the last two steps. We recommend reviewing [chapter 6](#) (to learn about identifying and collecting data) in conjunction with this chapter on design and methods. Selecting the appropriate designs and methods is an important part of evaluation planning because it helps you collect high-quality, relevant data to best answer your evaluation questions.

Evaluation terminology

In an evaluation, information is often referred to as data.

Evaluation plan templates

Many federal funding programs require grantees to develop and submit evaluation plans for agency approvals. Funders often have specific guidance for how grantees should develop their evaluation plans, but the general purpose is to make as many evaluation decisions as possible in advance of implementation to better ensure the timely and smooth execution of a program evaluation.

This Guide supports and aligns with common evaluation plan components, such as explaining evaluation questions, determining designs, identifying measures, and developing data collection procedures. For examples of evaluation plan advice or templates, see Children's Bureau, Administration for Children and Families (ACF), 2019, and Blocklin et al., 2019.

You may want to pull all your plans together into a formal evaluation plan. An evaluation plan is a "written document that describes how you will monitor and evaluate your program, as well as how you intend to use evaluation results for program improvement and decision-making" (CDC, 2011a). These plans are particularly helpful if membership in your evaluation team changes over time. The evaluation plan provides much information a new member will need to get oriented to the evaluation.

If made public, evaluation plans can also support transparency by sharing information with interested parties about how you plan to conduct your evaluation.

If you are conducting a high-profile evaluation, you may also use your evaluation plans to register your study (i.e., add your evaluation plan to a third-party study registry). Study registrations offer a more formal way to prespecify your evaluation approach and sometimes even your analysis plans. Registration brings transparency and credibility to your evaluation because you have registered a plan for how your study will unfold and what outcomes you will report. This disclosure helps prevent selective sharing of only positive findings in later reports.

Evaluation Design Versus Data Collection Methods

The distinction between evaluation design and data collection method is important. The two are different but closely related and easily confused. Evaluation design is the approach you use to answer each evaluation question. Evaluation data collection methods refer to the data collection strategies you use to execute that plan. No specific research design requires a specific data collection method. Therefore, you may select data collection methods after you select your design.

Because each program is unique, choose designs and data collection methods that fit your evaluation's goals, objectives, and expected ability to attribute outcomes as an effect of the program.

This chapter briefly introduces three types of evaluation designs (nonexperimental, quasi-experimental, and experimental) and describes the most frequently used data collection methods (surveys, secondary data analysis and archives, interviews, focus groups, and observations). These designs and data collection methods do not represent an exhaustive list but rather a starting point when considering the most suitable option for your program evaluation.

Treatment groups, comparison groups, and control groups

- **Treatment groups** are sets of individuals, classrooms, schools, departments, families, or other groups who are offered or receive the program being tested in the evaluation. In this Guide, we use the term treatment or program group to stay consistent with program as the focus of the evaluation.
- **Comparison groups** are sets of the same units as treatment or program who do NOT receive or are not offered the program. Comparison groups can be selected using numerous methods (see quasi-experimental designs below).
- **Control groups** are a specific type of comparison group created through random assignment. All control groups are comparison groups, but not all comparison groups are control groups. For simplicity, we use only the term comparison groups.

Refine Your Implementation Evaluation

Implementation evaluations can answer different types of questions, and each of those questions may call for different methods. Below are common implementation evaluation questions and their associated data needs (adapted from Blocklin et al., 2019):

- **Questions about reach.** Reach measures the scale of program activities. You could calculate reach at the participant level by counting the number of participants served by the program over a time period (such as yearly, or life of the program or evaluation). Community-level reach could be measured as the number of communities (or neighborhoods, schools, cities, or housing developments) served by the project.
- **Questions about saturation.** Saturation measures how widespread the program was. This concept is related to reach in that the count of people (or families, houses, schools, etc.) becomes a numerator. The denominator is a measure of the size of the overall population. You could calculate saturation as the number of children served over the number of children eligible for services or the number of children in the entire county, for example. Saturation is particularly important for programs that take a place-based or community-based approach to services.

- **Questions about service receipt.** Service receipt measures how participants engage in your program. This measure is more detailed than reach. It could measure the number and type of services participants (or different types of participants) received (e.g., 90 percent of men participated in job search activities, while 15 percent participated in substance prevention education). It can also measure dosage such as the percentage of people who completed the number of service hours required to graduate or the average number of mentor sessions youth attended.
- **Questions about fidelity.** Fidelity measures the extent to which the program was implemented as planned or as designed. Fidelity is measured separately for each program activity (e.g., outreach, recruitment, curriculum-driven sessions, case management meetings, coaching sessions). It is assessed by establishing a threshold for each activity (e.g., conduct 10 outreach sessions a month, recruit 200 families a year, 80 percent of enrolled participants receive at least 7 of 10 curriculum-driven sessions). Ideally, fidelity calculations occur at least annually as a way to inform program management decisions.
- **Questions about implementation drivers, barriers, and solutions.** These questions typically collect qualitative data from program staff, leadership, and partnering organizations to document how program implementation went. They focus on drivers (what helped the program function), barriers (challenges program staff encountered when running the program), and solutions (what program staff did to overcome or mitigate the barriers). Drivers and barriers can occur at many levels, such as federal, societal, state, local, system, agency, community, or individual.

Choose Your Evaluation Designs

Evaluation questions are often grouped in two categories: implementation (sometimes referred to as “process”) and outcome. Implementation evaluation questions are descriptive: They help you collect systematic information about how the program was delivered, who staffed the program, who participated in the program, how well program activities were delivered, and how external factors influenced program delivery.

Outcome evaluation questions document changes associated with the program, such as improvements in participant income, reductions in staff turnover, or changes in data interoperability. Typically, outcome evaluations are strengthened by accompanying implementation evaluations that provide context to outcome findings.

In designing evaluations to address outcome questions, evaluation teams must determine how to isolate the impact of a program from other factors that could influence the same outcome. Accordingly, outcome evaluation questions require a comparison condition; that is, a way to compare observed program results with those you would expect if the program had not been implemented

What are comparison conditions?

Outcome evaluations need to answer the question, compared to what? How will you know if the value of an outcome (e.g., average income after training) is “good?” You need to compare the outcome finding to something else. That something else is a comparison condition. Common comparison conditions follow:

- Pretreatment measures of the outcome from the treatment/program group
- A benchmark measurement such as state-level target for standardized test results
- An evidence-based target such as gains demonstrated by other similar program evaluations (e.g., a 10-percentage-point drop)
- Outcomes from a nontreated group, such as a comparison or control group, measured at the same time as the treatment/program group outcome

(i.e., the counterfactual¹). Evaluations establish a comparison condition for outcome evaluation questions via three designs: nonexperimental, quasi-experimental, and experimental.

These designs appear below in order from least to most able to attribute changes in outcomes to the program and not other factors:

Nonexperimental designs provide a hypothetical prediction of what would have happened in the absence of the program. The most common of these designs is a single group pre-posttest: Participants provide data on outcomes of interest, the program is implemented, and participants once again provide data on the same outcomes. For example, an evaluation captures participants' knowledge of child development before and after a parenting education program to examine whether participants demonstrate change (improvement) over time.

Other nonexperimental designs compare participant outcomes with benchmarks or national statistics (e.g., 85 percent of program children demonstrated grade-level reading skills compared with 70 percent of children of a similar age nationwide). Nonexperimental designs are often used when quasi-experimental or experimental designs are not feasible or practical.

Quasi-experimental designs identify a comparison group—individuals who are as similar as possible to the evaluation participants, but they did not participate in the program. Evaluators use many approaches to develop a quasi-experimental comparison group; for example, including—

- Individuals eligible for but uninterested in participating in the program
- Individuals not able to participate in the program yet because of program space constraints (wait list control group)
- Similar individuals in another community or school (matched comparisons; see textbox in [Choose Evaluation Samples](#) below)
- Artificial comparison groups created using advanced statistical techniques (such as synthetic comparison groups)

The evaluator can also use a design called comparative interrupted time series² to conduct quasi-experimental cluster evaluations, such as tests of community-, city-, or state-level interventions. Comparative interrupted time series use multiple data collection waves to establish patterns for both treatment/program and comparison clusters. Other advanced quasi-experimental designs make use of statistical techniques to develop a comparison condition, such as regression discontinuity or propensity score matching. To provide rigorous results, treatment groups and comparison groups must be statistically identical or similar to each other on pre-intervention measures of important outcomes.

¹ Counterfactuals allow evaluations to make comparisons between the observed results to those expected if the intervention had not been implemented (Better Evaluation, n.d.-a).

² Some evidence suggests comparative interrupted time series designs may be as internally valid as experimental designs (St. Clair et al., 2014)

Experimental designs also highlight comparison groups of people who were not offered the program. The difference is in how the comparison groups are constructed. In experimental designs, individuals are randomly assigned to be offered or not offered the program. Random assignment seeks to ensure the two groups are nearly identical in factors that may influence the outcome being examined. As a result, any difference in outcomes between the program participants and comparison group after the program has been implemented can be attributed to effects of the program.

Experimental designs offer the strongest evidence that changes in outcomes are caused by the program. Experimental designs are considered the gold standard in generating causal evidence and are important because they provide strong conclusions about whether a program should be replicated or expanded to more people, or whether the program is ineffective and should be discontinued or significantly revised.

Sometimes, randomly assigning some people to not be offered a program can seem unpalatable or objectionable to program staff or community members. Program staff may feel it is unfair to deny services to interested community members. It is important to remember the evaluation is providing information about whether a program works or even if it might have negative impacts. It is not unethical to offer a program to only a portion of interested people to learn whether it is working. This is similar to clinical trials in medicine (e.g., we know the polio vaccine is effective and so it is universally available, but we have not yet identified an effective vaccine against HIV, so clinical trials are ongoing).

If program staff or community members still have reservations about random assignment, evaluations may be able to overcome those reservations by addressing the services or programs offered to the comparison group. For example, not every evaluation needs to have a no-treatment comparison group, in which comparison group members receive no services. In some cases, you may be able to provide an alternate service or a small intervention such as a book or gift card. Alternatively, you could offer the program to comparison group members after final data collection ends.

Each of these designs offers different strengths and limitations (see table 5.1). Your evaluation team will need to select the most appropriate design for each outcome evaluation question based on factors such as the following:

- **Required level of rigor in causal attribution.** Your design choice will be influenced by the extent to which you must be able to document a change in an outcome has been caused by the program. The level of rigor you need may be dictated by your funder, your advisory board, or the goals of your evaluation (e.g., meet evidence requirements established by an evidence clearinghouse). In general,

Focusing on evaluation rigor

For each type of evaluation, several strategies can increase rigor or improve the quality of the information gathered. Examples follow:

- Nonexperimental designs can establish a priori goals for the magnitude of change expected to be seen if the program meets the level of improvement goals.
- Experimental and quasi-experimental evaluations can apply a “difference in differences” approach to calculate change over time, which helps to account for change that can be attributed to factors other than the program.

Work with your evaluator to identify feasible strategies to strengthen your evaluation designs.

experimental designs provide evaluators the highest confidence that any differences in outcomes are caused by the intervention. Quasi-experimental designs also enable evaluators to make inferences about whether interventions cause impacts, but there is some uncertainty about whether factors not observed by the study are causing changes in outcomes. Nonexperimental designs do not enable evaluators to demonstrate that changes in outcomes are caused by the intervention; these designs cannot rule out other factors as causing the changes. If you must have an evaluation at the highest level of rigor, experimental designs are most likely to meet that need.

- **Data availability.** Quasi-experimental and experimental designs need units of observation (e.g., families, centers, communities, children, classrooms) that did not participate in the program. That means you will need to identify comparison units and be able to collect identical data from both the comparison units and the units that were offered the program.
- **Resource availability.** Typically, more complex evaluation designs require additional resources. These resources include financial, technical skills/evaluation capacity, data capacities, and time to collect data on long-term outcomes (outcomes that happen months or years after the program ends). Some random assignment designs need a surplus of people interested in participating in the program to build large enough treatment/program and comparison groups to meet statistical power requirements.
- **Social, cultural, and political context.** Not all outcome evaluation designs are feasible in the “real” world. You will not be able to ask some families to not participate in a universal or mandated program (e.g., not attend public school). In other situations, community representatives may not approve of certain designs, particularly those that withhold an accessible program.

Table 5.1. Possible Designs for Outcome Evaluations

Design	Description	Example(s)	Use	Strengths	Limitations
Nonexperimental	<ul style="list-style-type: none"> ■ Designs without comparison groups or randomized assignment 	<ul style="list-style-type: none"> ■ Single group pre-post test 	<ul style="list-style-type: none"> ■ Describe individuals, settings, or events within the context of their occurrence 	<ul style="list-style-type: none"> ■ Can be used when baseline data and/or comparison groups are not available ■ Requires fewer resources 	<ul style="list-style-type: none"> ■ Minimal ability to infer causality
Quasi-experimental	<ul style="list-style-type: none"> ■ Designs with comparison groups but no randomized assignment 	<ul style="list-style-type: none"> ■ Matching and propensity score designs ■ Comparative interrupted time series designs ■ Regression discontinuity designs ■ Instrumental variables estimations 	<ul style="list-style-type: none"> ■ Conduct evaluations in field settings or in situations when comparable groups are created by differences that already occur in real world ■ More appropriate for complex community and systems change initiatives 	<ul style="list-style-type: none"> ■ Can infer moderate level of causality when not logistically feasible or not ethical to conduct randomized controlled trial 	<ul style="list-style-type: none"> ■ Offers moderate confidence in inferring causality ■ Differences between groups may generate a confound^a

Table 5.1. Possible Designs for Outcome Evaluations (continued)

Design	Description	Example(s)	Use	Strengths	Limitations
Experimental	<ul style="list-style-type: none"> ■ Designs with randomized assignment (inclusion of a control group) to definitively answer cause-effect questions 	<ul style="list-style-type: none"> ■ Randomized controlled trial 	<ul style="list-style-type: none"> ■ Establish cause-effect relationship ■ More appropriate for programs seeking highest level of rigor (considered gold standard for studying causal relationships) 	<ul style="list-style-type: none"> ■ Most robust design for testing causal hypotheses 	<ul style="list-style-type: none"> ■ Most resource intensive ■ Can be difficult to generalize to “real world”

Sources: CDC (2011b); Moore (2008)

^a Confounds are “any factor, other than the intervention, that is both plausibly related to the outcome measures and also completely or largely aligned with either the intervention group or the comparison group” (Wilson et al., 2019, p. 35). For example, if all treatment group members receive the treatment from a single individual, a confound is present because you cannot parse out whether the treatment (like a reading intervention) or the provider is responsible for the increase in reading scores in comparison to the comparison group.

Wait! Shouldn't we always conduct a randomized control trial?

Sometimes evaluators choose a design requiring random assignment because of its high prestige in the research community. Prestige, however, is not a relevant criterion for design selection. The best designs for your evaluation are those you can implement well and with design fidelity. Use the advice in this text box to determine whether a high-quality random assignment design is feasible in your situation. The evaluation literature describes common challenges to different evaluation designs; your evaluation team should critically evaluate their ability to overcome or ameliorate those challenges.

Many random assignment evaluations are underpowered: They are unable to recruit enough people in the treatment/program and comparison groups to actually detect differences in outcomes. What evidence do you have that you'll be able to recruit a sufficient number of evaluation participants?

Some programs undergo rigorous evaluation too soon—looking for program impacts before the program model has been refined and without any evidence the model is implemented with fidelity. This can make an otherwise effective program appear ineffective. Programs may want to conduct implementation evaluations first and test program improvements before turning to a rigorous outcome evaluation.

Regardless of which outcome evaluation designs you select, an implementation evaluation should ideally accompany an outcome evaluation. Pairing these two designs lets you provide context and strengthen interpretations of any outcome findings (e.g., an implementation evaluation can indicate if the program was poorly implemented to explain the lack of improvement in a related outcome).

Choose Evaluation Data Collection Methods

The next step in evaluation planning is to decide on the methods of data collection you will use to collect the information needed to answer your evaluation question. Many people think the term “data” refers to only numerical information, but data can be facts, statistics, images, quotes, or any other information collected about your program or participants.

Most evaluation questions can be answered using several data collection methods. Select those that best meet your needs, accessibility to various data sources, budget, and timeline. Common data collection methods include surveys, administrative data, interviews, focus groups, observations, and document reviews. The method you choose should be based on the type of data you want to collect (i.e., qualitative³ versus quantitative⁴). For example, focus groups, interviews, and observations are best for collecting qualitative data, while quantitative data are typically collected from survey and administrative data sources. However, most methods can produce both qualitative and quantitative data. For example, you can pose open-ended questions in a survey or provide percentages of the number of interviewees who identified the same implementation challenge in a focus group. See table 5.2 for a discussion of common data collection methods and their strengths and limitations.

Evaluation consent

For all methods that involve human subjects, it is imperative evaluations receive consent from people to collect data from or about them. See [Protect Study Participants](#) for more information.

Table 5.2. Possible Data Collection Methodologies

Design	Description	Example(s)	Use	Strengths	Limitations
Surveys	<ul style="list-style-type: none"> Data collection efforts that use a formal, prespecified, instrument to collect data; can be large-scale Can be paper and pen, online, phone fielded, or use combination of collection strategies Typically collect closed-ended questions (e.g., Do you like this Guide? Answer yes or no) but can also use open-ended questions (e.g., Describe how you will use this Guide) 	<ul style="list-style-type: none"> Online data collection form that evaluation participants complete 	<ul style="list-style-type: none"> Collect identical data across all evaluation participants; when you know all the questions you want to ask and what response options are 	<ul style="list-style-type: none"> Collect large amount of data from many people Produce quantitative data to inform statistical analyses 	<ul style="list-style-type: none"> Unlikely to collect new perspectives Can be costly to develop and field Need to invest in measurement selection and order of questions

³ Qualitative data are information that are difficult to measure, count, or express in numerical terms. For example, a participant's impression about the fairness of a program rule/requirement is qualitative data (OPRE, 2010).

⁴ Quantitative data are information that can be expressed in numerical terms, counted, or compared on a scale. For example, using a score developed from a reading test to document a child's reading level (OPRE, 2010).

Table 5.2. Possible Data Collection Methodologies (continued)

Design	Description	Example(s)	Use	Strengths	Limitations
Administrative data	<ul style="list-style-type: none"> Data that programs collect as part of providing services Can be data from your specific program, such as attendance records, but can also be data collected from your participants but by another organization 	<ul style="list-style-type: none"> Temporary Assistance for Needy Families recipient database 	<ul style="list-style-type: none"> Access data that have already been collected; reduce burden on evaluation participants; can be cost-effective 	<ul style="list-style-type: none"> Can be cost-effective; often very little missing data 	<ul style="list-style-type: none"> Measures of interest may not be available in existing datasets May be difficult to access based on data ownership May be difficult to link individuals in administrative data to other data sources May have challenges with data quality, accuracy, and/or thoroughness
Interviews	<ul style="list-style-type: none"> In-depth conversations with one or more individuals Interviewers typically use standardized protocols or lists of questions to guide the conversations Most questions are open ended 	<ul style="list-style-type: none"> In-depth, structured discussion with program manager 	<ul style="list-style-type: none"> Collect information about experiences, perceptions, or activities not easily captured with closed-ended questions 	<ul style="list-style-type: none"> Generate data when response options are unknown or too complex Able to shift and change as interviewee surfaces additional topics 	<ul style="list-style-type: none"> Resource intensive Requires significant time commitment on part of interviewee Data may be seen as less rigorous than other methods
Focus groups	<ul style="list-style-type: none"> Conversation held with multiple individuals at once Typically use written set of questions but also provide significant amount of space for focus group participants to react to, build on, and engage with comments from their fellow participants 	<ul style="list-style-type: none"> Focus group of program participants 	<ul style="list-style-type: none"> Collect data from multiple, similar evaluation participants at the same time Generate additional data through evaluation participant interaction and discussion 	<ul style="list-style-type: none"> Collect much data in short time May be more comfortable for evaluation participants than one-on-one interviews Can benefit from group synergy 	<ul style="list-style-type: none"> Can be difficult to schedule Group format may affect honesty of participant responses Need to be conducted by skilled facilitator Need to address issues of confidentiality with focus group participants
Observations	<ul style="list-style-type: none"> Members of data collection team “sit in on” event or process and document what they see Observations typically use tool to ensure consistent information is documented for each event 	<ul style="list-style-type: none"> Evaluator observes participant workshop session 	<ul style="list-style-type: none"> Document the “feel” of an event or process 	<ul style="list-style-type: none"> Give context and depth to an evaluation 	<ul style="list-style-type: none"> Typically need to be combined with other data source to answer evaluation questions

Table 5.2. Possible Data Collection Methodologies (continued)

Design	Description	Example(s)	Use	Strengths	Limitations
Document reviews	<ul style="list-style-type: none"> Make use of existing written materials as data sources Typically guided by extraction tools to help capture relevant information 	<ul style="list-style-type: none"> Program management decisions as documented in meeting notes 	<ul style="list-style-type: none"> Collect data already available in written sources 	<ul style="list-style-type: none"> Cost-efficient Less burdensome to evaluation participants 	<ul style="list-style-type: none"> Rely on quality, accuracy, and thoroughness of source documents

Typically, strong evaluations do not rely on just one type of data; instead, they employ a mixed-methods approach. Mixed-methods evaluations use more than one type of data to tell the program's story (i.e., they collect both qualitative and quantitative data). A specific kind of mixed-methods approach—triangulation—uses multiple methods to collect data on the same outcome. For example, an evaluation could triangulate customer satisfaction by surveying participants (quantitative), interviewing participants (qualitative), and conducting observations of participant and program staff interactions (qualitative).

Protect Evaluation Participants

Rapid cycle evaluation (RCE)

RCE is an approach to evaluation that relies on innovative design and methods to quickly test program components and provide actionable results to integrate improvements into further testing. With RCE, program or process changes can be tested in a shorter time and decision-makers can have increased confidence in results. For a more detailed look at RCE, see Atukpawu-Tipton and Poes (2020).

All evaluation efforts, including data collection, must respect and protect the privacy of the individuals who contribute information to the evaluation. Evaluators and social science researchers have developed procedures designed to ensure individuals who provide data do so voluntarily, have their information safeguarded, and have their privacy respected.

Institutional review boards (IRBs) are oversight agencies that review study procedures to ensure study participants' rights and welfare are protected. Many large evaluation

firms, almost all universities, and numerous state-level agencies have IRBs that approve evaluations conducted by their staff or with their funding. Independent evaluators can hire private IRBs to approve their evaluations. Any human subjects research or evaluation conducted with federal dollars is required to receive IRB approval.

Broadly speaking, protecting study participants (i.e., any human who provides information to a study, including evaluation) includes the following:

- **Informed consent.** People who contribute information about themselves for an evaluation should understand what information you are asking for, what you are doing with the information, how you will protect their information and identity, what risks there may be if they participate in the evaluation,

and what would happen if they choose to not participate in the evaluation (e.g., they would still receive services but wouldn't receive financial incentives for data collection). Consent procedures should take into account participants' preferred languages, levels of literacy, comfort with research, and power dynamics between the evaluation team and potential participants.

- **Voluntary nature.** Individuals must retain their rights to autonomy. Potential evaluation participants should not be coerced into participation in the evaluation, nor should they face significant consequences for not participating in the evaluation. Evaluation participants should be able to refuse to answer questions or provide information they are not comfortable providing. Evaluation participants should also be able to revoke their consent at any time. Evaluations should provide contact information if an evaluation participant wants to have their data removed from the evaluation and destroyed at a later point.
- **Data security procedures.** Your evaluation team will need to develop procedures and safeguards to ensure only the people who need to see the data (i.e., evaluation team members) can see the data. This includes safeguarding paper copies of surveys, online databases, participant contact information, and datafiles.
- **Privacy procedures.** Your evaluation team will also need to determine if, and if so how, you will ensure

information provided by individuals won't be linked to them. Such procedures include maintaining participant contact information separate from other datafiles, using unique identifier codes rather than people's names or Social Security numbers, and having reporting standards around cell sizes.

Matching comparison and treatment/program groups

When choosing a sample for a quasi-experimental evaluation, you need to consider how to develop a comparison group that is as similar as possible to your treatment/program group. These similarities reduce the strength and number of alternate explanations for differences in outcomes between the two groups.

This means you will want to select demographic characteristics (e.g., age, gender, race/ethnicity), characteristics associated with your outcomes of interest (e.g., programs aiming to improve educational attainment should choose a comparison group with a similar mixture of educational credentials to the treatment/program group), and temporality (e.g., avoid measuring treatment/program group wages in 2018 and 2022 and using a comparison group with wage data from 2004 and 2008).

Confidentiality and privacy

What's the difference? Confidentiality and privacy are both important concepts in protecting evaluation participants' data, but they have different meanings. Privacy is about people, while confidentiality refers to data. Evaluations protect participants' privacy by collecting only information the evaluation needs and using discreet data collection procedures (e.g., interviews in a private room, not in a public cafeteria). Confidentiality extends privacy by protecting the information participants provide. It includes procedures used to ensure only authorized people have access to data, and they won't purposely or inadvertently identify evaluation participants and share their information (UCI Office of Research, 2021).

Choose Evaluation Samples

For each data collection method described above, you will need to determine the data sources for the evaluation. Two terms evaluators use when talking about how units of observation can be selected follow:

- **Census.** A census means you collect data from each unit of observation eligible to provide data. For example, you may survey every single individual who participated in a program (and consents to data collection).

- **Sample.** A sample means you select some number of the eligible units of observation to provide data. For example, your evaluator will not be able to observe every single interaction between case managers and program participants. In that case, you and your evaluator will need to determine how you will develop your sample.

Each different technique for sampling has different implications for your evaluation budget, evaluation timeline, and the extent to which your data can be generalized or seen as representative of the whole group of units of observation. The higher the generalizability, the more likely the data collected from the sample are similar to data you would have gotten if you had collected data from every member of that group. Below are a few common approaches to sampling:

- **Random sampling** means everyone has the same probability of being chosen to be a part of the sample. For example, if you need to collect neighborhood data through a survey, you could knock on every 10th door in the community. Random sampling has strong generalizability and can be cost-effective by helping save time and resources.
- **Convenience sampling** collects data from units of observation easiest to reach. If you recruited for a staff focus group by emailing a request for volunteers, you would have a convenience sample. Convenience samples are easier to generate because you know participants are interested in engaging in the evaluation. However, they provide poor generalizability because it's hard to know if eager individuals differ in important ways from people who didn't see or answer a call for engagement.
- **Purposeful sampling** takes into consideration the purpose of the evaluation, along with the understanding of the target audience. For example, for a purposeful interview sampling frame, your evaluator may call five people who never completed their intake, five people who attended only the first session of the program, and five people who completed the whole program. This strategy would enable the evaluator to answer questions about the program from numerous perspectives and level of engagement.

Sampling approaches can be challenging to develop, implement, and document. If you think your evaluation will need to collect data from a sample of units of observation, it is important a member of your evaluation team has sampling experience.

Manage and Monitor the Evaluation

As part of designing your evaluation, you will need to build in systems and time to track and manage the evaluation, including data collection and program staff engagement in the evaluation and procedures for updating and revising your evaluation as needed.

Common strategies and tools to manage an evaluation follow:

- **A written evaluation plan.** Drawing from the work you do to develop the evaluation, consider formalizing the final decisions in a written evaluation plan. See the end of the chapter ([To learn more...](#)) for resources with guidance on developing evaluation plans. Most include evaluation questions, designs, data collection methodologies, measures, analysis plans, roles and responsibilities, and an evaluation timeline.
- **Staff trainings and manuals.** Often program staff are engaged in aspects of evaluation data collection, such as fielding intake forms or documenting attendance at events. Everyone engaged in the evaluation must understand how to support the evaluation correctly. This will ensure consistency in information collection and be useful for staff who are hired after the evaluation begins. Training materials help explain the purposes of the evaluation and data collection, how to collect and input data, common challenges to accurate data collection, and advice for solving common challenges.
- **Data dictionaries and coding manuals.** Evaluation team staff engaged in assessing and analyzing data should follow standardized practices to code, clean, and analyze data. Manuals and guides, typically developed by the lead evaluator if delegating work to others, can help support rigorous data analysis.
- **Data quality monitoring.** Evaluation team members should regularly check all data collected to ensure forms are completed accurately, identification numbers are used correctly, and no more than an acceptable amount of data is missing (the amount of acceptable missing data will depend on your study type, expected level of rigor, and funder guidance).
- **Continuous quality improvement procedures.** Programs often engage in continuous quality improvement efforts to identify ways to strengthen program implementation or management. Evaluations may want to apply those same concepts. For example, conversations with data collection staff can help identify if any procedures are hard to follow or burdensome.
- **Regular evaluation team meetings.** Check timelines, review current work, identify, and navigate challenges.

Finally, evaluations often have yearlong timelines or timelines across many years. It is unrealistic to expect all program operations to stay static over that time. For example, you may add or discontinue a particular service or program component. Your evaluation should keep track of changes in program operations through procedures for documenting the time this change occurred, the reasons for the change, and whether particular participants were engaged in the program prior to or after the change. This will help you determine whether the change had any impact on attainment of expected outcomes.

Navigate Unexpected Changes

Over the course of your evaluation, you likely will not implement the evaluation exactly the way you intended. Changes are expected; you might even make changes to improve the quality or rigor of your evaluation. Examples of possible evaluation changes include program changes, changes in funding, staff turnover, losing access to a data source, slipping timeline, or changes in your expected sample size.

Conducting evaluations during a pandemic

COVID-19 upended much of daily life, program evaluations included. As programs cancelled in-person activities and pivoted to virtual services, program evaluations also needed to adapt to new public health restrictions. Evaluators revisited their evaluations and worked with programs to determine which evaluation questions were still applicable, which data collection efforts could be adapted to virtual efforts, and how to consent participants and protect their data in a virtual environment.

Evaluators who successfully mitigated the setbacks and challenges posed by COVID-19 worked as partners with their program staff. They brainstormed innovative ideas and pressure-tested them for feasibility. They looked to colleagues and other fields for advice and support. Many evaluators also introduced new evaluation questions to understand and document how programs adapted and responded to the massive changes brought about by COVID-19.

When changes occur, you should take the following steps:

Step 1

Determine how the changes will affect the evaluation

For example, if you will end up with a smaller sample size, your study may have less statistical power to detect changes in outcomes. If you lose access to a data source, you may no longer use those data to measure program outcomes.

Step 2

Determine whether those effects are acceptable, and if not, develop an alternative plan

Continuing with our examples, your statistical expert may indicate your expected magnitude of change can still be detected with a smaller sample size. Conversely, you may determine the outcome measure collected by the lost data source is key to your evaluation and develop another method for collecting similar information.

Internal and external validity: What are they and why do they matter?

Evaluators and researchers use these terms to assess certain elements of credibility of studies. Internal validity refers to the extent to which an evaluation identifies the true impact for the individuals included in the evaluation. Many evaluation decisions or quality of execution can affect internal validity. For example, random assignment to treatment/program and comparison groups and high response rates for data collection efforts increase internal validity. Differential attrition (where individuals in the treatment/program and comparison groups drop out of the study at different rates), crossover (where comparison group members receive the treatment/program), and having program staff collect data (as opposed to an independent, “unbiased” data collector) can threaten internal validity.

External validity refers to the extent to which the results of an evaluation might be applicable in other situations (also known as generalizability). Both program and study elements can affect external validity. For example, large incentives or program supports that aren’t feasible in the “real world” reduce external validity, while implementing the program in a standard setting, as opposed to a laboratory or clinical setting, increases external validity.

Step 3

Document the change, and if needed, the solution

Good evaluators develop sufficient documentation of their evaluation to enable them to answer questions about the process and potentially for another evaluator to replicate the evaluation. You should document when the changes occurred (potentially even coding your dataset like a variable for whether participants received the initial or revised program) and how you handled them. This information is also good “institutional knowledge.” Evaluation teams sometimes experience turnover, and strong documentation can help a new team member quickly get up to speed.

Step 4

Prepare and plan for changes in the future

While this Guide may support you through one evaluation, you will likely conduct others in the future. Be sure to draw from your experience with changes and challenges during your current evaluation and proactively plan for them. For example, if you had sample size issues, in future evaluations you may want to plan for a much larger than needed sample (by recruiting more people to participate in a program, or by enrolling people in the evaluation over a longer time period).

Practice Culturally Responsive and Equitable Evaluation When Designing an Evaluation

A CREE approach to selecting your evaluation designs, data collection methodologies, and sampling frames requires a critical dialogue with community members. Engaging with community representatives will help ensure the evaluation is co-created and community members have buy-in and view the evaluation as credible and meaningful.

Evaluation designs. Reflect on whether, or to what extent, certain evaluation designs may be a poor match for the program participants and community members. For quasi-experimental and randomized control groups, can you match treatment and comparison groups on similar characteristics determined as noteworthy, such as race? Also consider the implications or appropriateness for assigning potential participants to a comparison group. For example, communities with a history of purposeful exclusion from beneficial programs and policies, such as difficulties accessing the GI bill for Black veterans (Smithsonian American Art Museum, n.d.), may have particular difficulty accepting a random assignment model where the comparison group is offered no or few services. It may be possible to balance evaluation information needs with community needs. For example, the comparison group of an evaluation of a job training program could be offered weekly social support groups rather than no services at all, or the comparison group could be offered services after all waves of data collection.

Data collection methods. Have conversations with evaluation users and community members about what “counts” as credible evidence. Often evaluations tend to prefer quantitative data and use qualitative data to supplement or add context to quantitative results. Data collection methods that capture more individualized experiences (e.g., interviews, focus groups, photovoice, appreciative inquiry, ripple effects mapping) could resonate better with evaluation users and help the evaluation team develop a more complete understanding of possible implementation and outcome findings.

Survey tools and sample sizes. Conventional evaluation encourages the use of standardized measures and instruments. These measures may not have been validated with people who would respond or interpret questions similarly to your evaluation respondents. They might also call for larger sample sizes for generalizability than you could reach if also disaggregating data based on demographic characteristics, such as race and religion. That doesn't mean you should or shouldn't use them. Just be aware of the considerations for each choice. Ideally, you can use a mixed-methods approach where some data sources could lead to generalizable findings, while others might need to be considered within the context of the group where they were collected.

Protecting evaluation participants. If a CREE approach has been incorporated into your evaluation design, consider how it could influence the steps needed to protect evaluation participants. For example, if using data collection methods that have participants share their opinions in front of one another, that approach changes what confidentiality and privacy entail. Each participant must commit to confidentiality and trust in one another other to carry through. Although not a data collection topic, this could also apply to community members who are on the evaluation team or an advisory board. Consider if the level of vulnerability or trust you are requesting is necessary and appropriate.

Data collection protocol. It is important to establish rapport before expecting participants to share their experience, story, or personal or private information. Build in time before data collection for data collectors to connect with evaluation participants on a personal yet professional level. Discuss aspects of privacy and confidentiality; if there are multiple participants in the room, also discuss considerations for creating a safe space for sharing.

What is credible evidence?

Consider the following example of determining what “counts” as credible evidence. A program helping parents ensure their children develop in an enriching, stable home environment is conducting an outcome evaluation of their efforts. They know they need to measure financial outcomes for families. The evaluator initially recommends they measure whether parents were able to buy a vehicle for child transportation.

During discussions with former program participants, the evaluators hear that many parents don't want a car, per se. Instead, they felt a more accurate marker of stability was whether they were able to transport their family to places they needed to be, regardless of transportation modality (e.g., rideshare, bus, bicycling, car rental). If you measure only car ownership, your evaluation will miss the evidence of impacts the community wants to see through program participation. In that sense, the evaluation will not generate evidence that is of interest to, or credible to, the community.

To learn more ...

- [Building Strong Evidence in Challenging Contexts: Alternatives to Traditional Randomized Controlled Trials](#) (Malin & Deterding, 2017)
- [Evaluation Plan Template](#) (CDC, n.d.)
- [Examining the Internal Validity and Statistical Precision of the Comparative Interrupted Time Series Design by Comparison With a Randomized Experiment](#) (St. Clair et al., 2014)
- [Manager's Guide to Evaluation](#) (Better Evaluation, n.d.-b)
- [Quantitative Research Designs: Experimental, Quasi-Experimental and Descriptive](#) (Drummond & Murphy-Reyes, 2018)
- [Quick Guide to Sampling, Sample Sizes, and Representation](#) (Washington State University, 2020)
- [Sampling and Evaluation: A Guide to Sampling for Program Impact Evaluation](#) (Lance & Hattori, 2016)

References

- Atukpawu-Tipton, G., & Poes, M. (2020). *Rapid cycle evaluation at a glance* (OPRE Report 2020-152). U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation. <https://www.acf.hhs.gov/opre/report/rapid-cycle-evaluation-glance>
- Better Evaluation. (n.d.-a). *Compare results to the counterfactual*. <https://www.betterevaluation.org/en/rainbow-framework/understand-causes/compare-results-to-counterfactual>
- Better Evaluation. (n.d.-b). *Manager's guide to evaluation*. <https://www.betterevaluation.org/managers-guide>
- Blocklin, M., Hyra, A., Kean, E., & Porowski, A. (2019). *Building capacity to evaluate child welfare community collaborations to strengthen and preserve families (CWCC) grantee local evaluation plan and implementation plan templates*. Abt Associates. <https://omb.report/icr/201906-0970-001/doc/98252801.pdf>
- CDC (Centers for Disease Control and Prevention). (n.d.). *Evaluation plan template*. https://www.cdc.gov/tb/programs/Evaluation/Guide/PDF/Evaluation_plan_template.pdf
- CDC. (2011a). *Developing an effective evaluation plan: Setting the course for effective program evaluation*. <https://www.cdc.gov/obesity/downloads/cdc-evaluation-workbook-508.pdf>
- CDC. (2011b). *Introduction to program evaluation for public health programs: A self-study guide*. U.S. Department of Health and Human Services. <https://www.cdc.gov/evaluation/guide/CDCEvalManual.pdf>
- Children's Bureau, ACF (Administration for Children and Families). (2019). *Evaluation plan development tip sheet*. U.S. Department of Health and Human Services. <https://www.acf.hhs.gov/cb/policy-guidance/im-19-04>
- Drummond, K. E., & Murphy-Reyes, A. (2018). *Quantitative research designs: Experimental, quasi-experimental, and descriptive*. In *Nutrition research: Concepts and applications*. http://samples.jbpub.com/9781284101539/9781284101539_CH06_Drummond.pdf
- Lance, P., & Hattori, A. (2016). *Sampling and evaluation: A guide to sampling for program impact evaluation*. https://www.researchgate.net/publication/311805268_Sampling_and_Evaluation_A_Guide_to_Sampling_for_Program_Impact_Evaluation/citations
- Malin, J., & Deterding, N. (2017). *Building strong evidence in challenging contexts: Alternatives to traditional randomized controlled trials*. U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation. https://www.acf.hhs.gov/sites/default/files/documents/opre/methodsmeetingsummary2016_final_112017_508.pdf

- Moore, K. A. (2008). *Quasi-experimental evaluation: Part 6 in a series on practical evaluation methods* (Publication 2008-040). Child Trends. https://www.childtrends.org/wp-content/uploads/2008/01/Child_Trends-2008_01_16_Evaluation6.pdf
- OPRE (Office of Planning, Research, and Evaluation). (2010). *The program manager's guide to evaluation, Second Edition*. U.S. Department of Health and Human Services, Administration for Children and Families. <https://www.acf.hhs.gov/opre/report/program-managers-guide-evaluation-second-edition>
- Smithsonian American Art Museum. (n.d.). *After the war: Blacks and the G.I. Bill*. <https://americanexperience.si.edu/wp-content/uploads/2015/02/After-the-War-Blacks-and-the-GI-Bill.pdf>
- St. Clair, T., Cook, T. D., & Hallberg, K. (2014). Examining the internal validity and statistical precision of the comparative interrupted time series design by comparison with a randomized experiment. *American Journal of Evaluation*, 35(3), 311–327. <https://doi.org/10.1177/1098214014527337>
- UCI Office of Research. (2022). *Privacy and confidentiality*. <https://research.uci.edu/human-research-protections/research-subjects/privacy-and-confidentiality/>
- Washington State University. (2020). *Quick guide to sampling, sample sizes, and representation*. <https://ace.wsu.edu/documents/2015/03/sample-size-and-represent>
- Wilson, S. J., Price, C. S., Kerns, S. E. U., Dastrup, S. D., & Brown, S. R. (2019). *Title IV-E Prevention Services Clearinghouse Handbook of Standards and Procedures, version 1.0* (OPRE Report 2019-56). U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation. https://www.acf.hhs.gov/sites/default/files/documents/opre/psc_handbook_v1_final_508_compliant.pdf