

Chapter 7. Analyze Data

What's **Inside?** _____



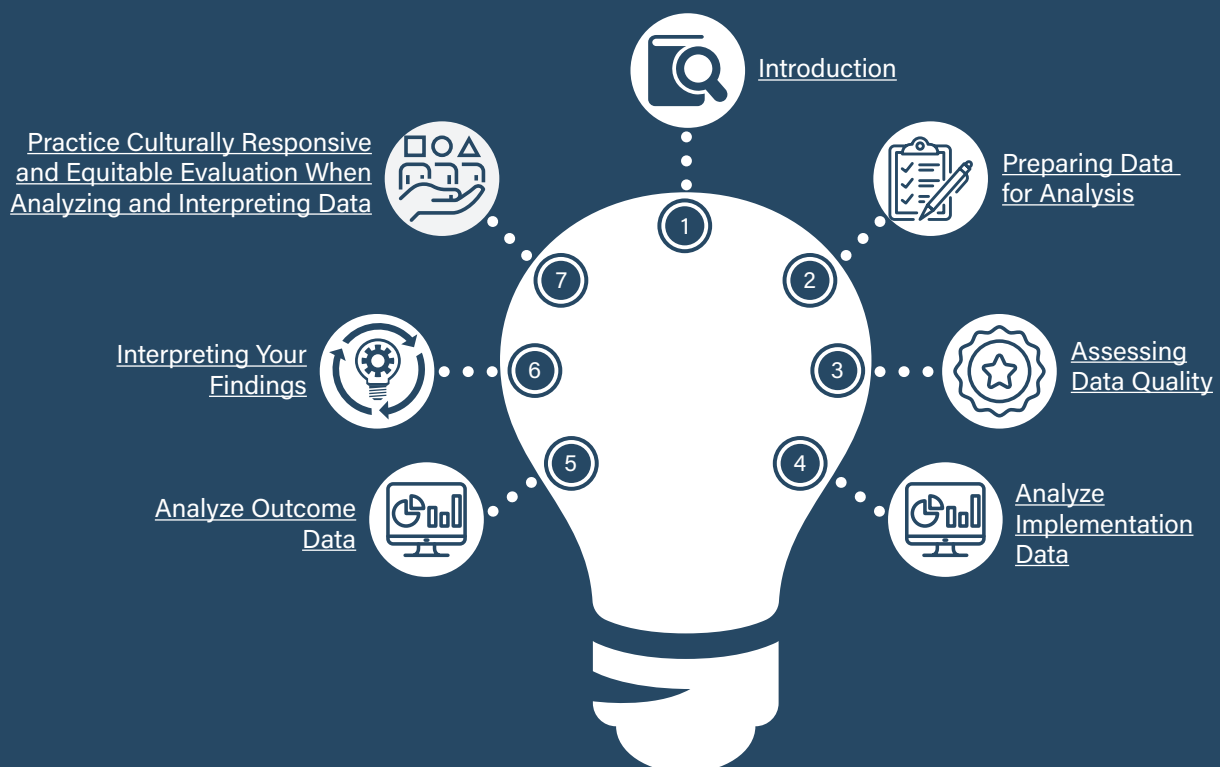
What this chapter contains

- An introduction to analyzing collected data and interpreting what the data (findings) mean
- A description of common procedures for preparing data for analysis
- Recommendations for assessing data quality
- A discussion of procedures for analyzing implementation and outcome data
- Examples of ways to apply culturally responsive and equitable principles when analyzing and interpreting evaluation data

Who can use this chapter

- Program managers preparing to analyze and interpret evaluation data

Click the links below to view the relevant section



Introduction

After you have gathered evaluation data, the next stage is to analyze those data and interpret what the findings mean. Although someone experienced in working with quantitative and/or qualitative data should lead data analysis, all members of the evaluation team should be engaged in making decisions throughout this stage. You will need to make decisions related to how your evaluation team will—

- Prepare the data for analysis (data cleaning and transformation).
- Conduct an initial assessment of data quality.
- Conduct analyses needed to answer evaluation questions.
- Potentially, conduct additional exploratory analyses.
- Discuss initial findings with community representatives to ensure their perspectives inform your interpretations.
- Interpret findings and make meaning.

While this chapter is not a manual for conducting statistical tests to analyze evaluation data, it provides basic information about approaches to analyzing evaluation data to help you understand and participate more fully in this process. Many ways to analyze and interpret evaluation data are available, and the methods discussed in this chapter are not the only possibilities. Whatever methods your evaluation team decides on, be sure your evaluation questions guide your analysis. The following evaluation questions are discussed throughout this manual:

- **Has the program been successful in attaining the anticipated implementation objectives?** If not, why not? What types of barriers impeded implementation objectives, or what factors facilitated their attainment?
- **Has the program been successful in attaining the anticipated outcomes?** If not, why not? What types of barriers impeded outcome objectives, or what factors facilitated their attainment?

The following sections discuss various approaches to analyzing data to answer both types of questions. The chapter concludes with examples of how to apply culturally responsive and equitable principles when analyzing and interpreting evaluation data.

Preparing Data for Analysis

Data are not ready to be “crunched” or analyzed right after collection. Both qualitative and quantitative data need to be prepared.

Clean the data. Cleaning data involves examining your data to ensure accuracy and completeness. Accurate data do not have any remaining incorrect or erroneous values for any element. For example, if you

have seven racial/ethnic codes with values of 1–7, you should not have any individuals with an out-of-range value of 12 or -6. You might explore whether some variables have values that seem improbable, such as birth dates in the 19th century or a parent indicating they have 26 children.

Ensure the type of data is correct. For example, numeric data elements should not have any character responses, such as a response of "&" when you expect a value between 0 and 3. Invalid or unlikely values may indicate an error was made in entering the data. You and your team will need to decide which values seem invalid and how you will handle those data, such as recoding erroneous values as missing.

Assess data accuracy. Develop an approach to assess accuracy of the data. For qualitative data, you could share interview notes back with the interviewee to see whether you captured the conversation correctly; compare notes with transcripts; or if you had more than one interviewer, have them compare and reconcile notes. For quantitative data, you might verify agreement on responses to related items. For example, if a survey respondent indicates in one response that they live with their spouse but in another says their household size is one, one response is likely incorrect.

Transform the data. Another common preparation step is to transform data. This task is particularly common with quantitative data, and it will depend on your evaluation questions and approach to the work. One transformation is the calculation of a scale score. Standardized instruments often come with guidance about how to convert responses into a numerical score. You may also want to collapse categories; for

example, recoding the number of children to 0, 1, 2–3, and 4 or more (if you have indications those differences matter).

Research and evaluation clearinghouses

Clearinghouses are repositories of high-quality program evaluations that try to answer questions of effectiveness. The federal government has funded numerous evidence clearinghouses such as the What Works Clearinghouse (Department of Education, n.d.), CLEAR (U.S. Department of Labor, n.d.), the Prevention Services Clearinghouse (ACF [Administration for Children and Families], n.d.-a) and the Home Visiting Evidence of Effectiveness project (ACF, n.d.-b). Most of these clearinghouses examine data quality issues such as study attrition in their review of impact evaluations. It is beneficial to use information about clearinghouse standards to guide the review of your impact evaluation (quasi-experimental or experimental) data.

Analysis is an ongoing activity

For many reasons, you should not wait until 6 months before your final report is due to start conducting analysis. Interim analysis conducted regularly throughout the course of your program evaluation (1) helps identify any major challenges to the data and facilitates midcourse corrections, (2) informs continuous quality improvement efforts by identifying challenges with program implementation, and (3) gives a preview when changes in outcomes are generally proceeding the way you expect (even if your interim samples are too small to detect statistical significance). Consider conducting annual analysis efforts.

Link the data. Often, evaluations collect many waves of data that must be combined for analyses. Studies need to match respondents across datafiles using identifying information. Ideally, each individual has a unique identifier value to combine datasets. If you don't have an identification number, or the number isn't applied accurately across all cases, you might be able to use techniques such as probabilistic matching to make likely merges of same-respondent data across different files (see Asher et al., 2020, for an introduction). If you merge data to create an analytic file, examine your new datafile to ensure accurate merging.

Transcribe the data. In the case of qualitative approaches, you may need to transcribe and clean interview or focus group data. Transcription is the process of converting speech (either live or recorded) into a written or electronic text document to facilitate coding of qualitative data. While transcribing may appear to be a straightforward technical task, the process of transcription may differ according to its end use (see Bailey, 2008, for more guidance).

Memoing

Memoing refers to informal record-keeping by qualitative researchers that references ideas, hypotheses, research literature, or observations about evaluation questions, design, methods, and theory as they arise throughout the evaluation process (see Satterlund, n.d.). Memoing helps you keep track of your thoughts and support evaluation team communication.

Decisions about the level of detail (e.g., whether to transcribe or omit nonverbal communication) should be discussed in advance. For example, if you plan to use transcripts to identify quotes and sound bites for supporting evidence, you may not need the same level of detail as for those that will be systematically reviewed, grouped into themes, and analyzed for content.

When transcripts are completed, the evaluation team should engage in data familiarization,¹ a common practice in all forms of qualitative data analysis. Researchers may begin identifying and notating features of the data that are potentially relevant to the evaluation questions, a helpful step in preparing for the next phase, the coding process.

Types of transcripts

Several kinds of transcripts can be used in qualitative research, depending on the methodology used and purpose of an evaluation:

- *Verbatim transcripts* are the most common type of transcript used for thematic analysis; they aim to capture every word and nonverbal auditory communication (e.g., sighs, laughing, stutters, pauses).
- *Jeffersonian transcripts* are designed to capture what was said and the way it was said, using symbols to represent sound, pace, intonation, and interaction in the conversation.
- *Gisted transcripts* are less detailed than verbatim or Jeffersonian transcripts; they aim to capture the essence (i.e., “the gist”) of an audiofile or videofile’s content.
- *Multimodal transcripts* are commonly used when analyzing video recordings of interviews, focus groups, or other forms of social interaction. All nonverbal forms of communication (e.g., gaze, head shake, gestures, eye rolls, posture) and verbatim communication are transcribed. Other variables that may influence participant responses are noted (e.g., cell phone ringing) to produce a highly comprehensive set of analyzable data.

¹ Data familiarization is the process of repeatedly reading or listening to each item of data to develop a deeper understanding of participants’ perspectives (Statistics Solution, n.d.).

Assessing Data Quality

Before you invest great effort analyzing data, understand the quality of data that will generate your findings. You cannot create better data through analysis.

One of the most important markers of data quality is the extent of missing data. Calculate the response rate for data collection efforts, such as the percentage of evaluation participants who completed a customer satisfaction form over the total number of evaluation participants asked to complete the form. Calculate the rate of missing data for each item also. For example, your overall survey may have had a high response rate, but one item about the amount of meat eaten per week had much missing data.

Study attrition. You must also calculate study attrition.² For example, if 100 people completed the pretest, how many of them completed the posttest? If your evaluation used a comparison group, calculate two types of attrition for each data collection wave: overall attrition, for the whole sample, and differential attrition, the difference in attrition rates from the treatment versus comparison group. High levels of differential attrition can indicate your two groups are too different to produce reliable comparisons that could be attributed to the program being evaluated. Attrition is calculated on both the sample and the measure level. In other words, report the number and percentage of your sample that provided any data at each wave of data collection. Also report the number and percentage of your sample that provided data for each outcome measure at each wave of data collection. This means that measures with more missing data (as described above) have higher attrition rates than measures from the same instrument and data collection wave with fewer missing responses.

Measurement reliability. You may want to assess the reliability of some of your measures (see [chapter 6](#) for more information). If you have developed your own scales or are using a scale not yet validated with your population, check whether it seems to have captured the construct. One approach to this task is to calculate a Cronbach's alpha statistic (UCLA, n.d.), which is a measure of how well all the items in your scale relate to one another (i.e., together capture the same construct).

Similarly, if you have a measure captured by observation (e.g., data collectors rated child behavior in the classroom), consider calculating its interrater reliability.³ To do this, you will need more than one data collector to collect the same observation data on the same evaluation participants. These analyses will demonstrate the level of agreement between data collectors. High agreement means the observation tool is reliable regardless of which data collector is completing it.

Qualitative data quality. When conducting qualitative research, data collection often runs concurrently with data analysis, and a high level of rigor in qualitative data is often discussed as the level of trustworthiness. Trustworthiness is established when findings as closely as possible reflect the meanings as described by

² Study attrition refers to any loss in responses from the study sample (Deke et al., 2015).

³ Interrater reliability is the degree to which different raters or judges make consistent estimates of the same phenomenon; also known as interobserver reliability (Multon & Coleman, 2018).

the evaluation study respondents. The researcher is often the primary instrument for data collection in qualitative approaches, so researcher biases not adequately addressed or errors in judgement can affect the quality of data and subsequent interpretation of findings. Unlike quantitative methods' strong emphasis on reliability⁴ and validity,⁵ it is not possible to use the same metrics when judging the quality of conclusions in qualitative studies. Instead, more viable alternatives have been proposed to serve as evaluative criteria (see table 7.1).

Table 7.1. Terms Used to Establish Trustworthiness of Qualitative Data

Conventional Terms in Quantitative Research	Alternative Evaluative Criteria in Qualitative Research	Description of Criteria for Demonstrating Rigor in Qualitative Research
Objectivity	Confirmability	Requiring researchers to be reflexive or self-critical about their biases
Internal validity	Credibility, authenticity	Presenting an accurate description or interpretation of a human experience
External validity	Transferability, fittingness	Transferring findings or methods from one group (or setting) to another
Reliability	Dependability, audibility	Following the decision chain, so other researchers can determine the credibility of the findings

Several strategies are available to evaluation teams to establish trustworthiness of qualitative data. Throughout the research process, the evaluation team should practice reflexivity.⁶ Similarly, the use of an audit trail⁷ offers flexibility to make decisions not previously prescribed while still requiring justification of those decisions to be recorded.

Triangulation, peer debriefing, and member checking can avoid or minimize error or bias and boost the accuracy in data collection and analysis processes:

- Triangulation involves identifying convergence of data obtained through multiple data sources and methods (e.g., observation field notes and interview transcripts).
- Peer debriefing, sometimes referred to as analytic triangulation, involves consulting with researchers outside an evaluation project who have experience with the topic, the population, or methods being used to better explain how the evaluation team's own values and interests are influencing the conduct, interpretation, and analysis of the research project. Peer debriefing is often compared with internal validity.

⁴ Reliability refers to the extent to which a measurement (such as an instrument or a data collection procedure) produces consistent results over repeated observations or administrations of the instrument under the same conditions. It is important that reliability be maintained across data collectors; this is called interrater reliability (OPRE, 2010).

⁵ Validity refers to the extent to which a measurement instrument or test accurately measures what it is supposed to measure. For example, a reading test is a valid measure of reading skills but is not a valid measure of total language competency (OPRE, 2010).

⁶ Reflexivity refers to actively acknowledging how one's own identity, beliefs, and values are inevitably assisting or hindering the process of co-constructing the meaning of the experience under investigation (University of Melbourne, n.d.).

⁷ An audit trail refers to clear documentation of research procedures throughout the data analysis process (Robert Wood Johnson Foundation, 2006).

- Member checking refers to a set of processes evaluation teams can use to “check in” on how participants in qualitative data collection respond to comments in the data or to researchers’ interpretations of the data. Ideally, member checks are used in combination with other methods to establish a study’s credibility (i.e., ensure the research findings are believable to participants).

Document all data quality issues in your reports, and discuss the implications and limitations associated with your findings based on data quality. In some cases, your data quality issues may be so severe you cannot use some elements or even an entire dataset.

Analyze Implementation Data

As a reminder (see [chapter 1](#)), implementation evaluations use data on program implementation to assess whether and to what extent program activities are being implemented as planned, expected program services are being delivered as planned, and how the program is operating in practice. Examples of basic program implementation evaluation questions follow:

- **How will we know the planned activities occurred?** For example, the number, duration, and frequency of services or activities implemented
- **Who will do it?** What the staffing arrangements will be; the characteristics and qualifications of the program staff who will deliver the services, conduct the training, or develop the products; and how these individuals will be recruited and hired
- **What population do you plan to reach, and how many individuals?** A description of the participant population for the program, the number of participants to be reached during a specific timeframe, and how you plan to recruit or reach the participants

Implementation evaluations typically collect data about implementation barriers and facilitators and how staff and program participants experienced the program. Because implementation evaluations do not try to ascribe changes to the program, they rely on descriptive analyses. Descriptive analyses paint a picture of the setting and provide details but do not attempt to measure the association or relationship between measures. Descriptive analytical techniques can be applied to both qualitative and quantitative data.

Quantitative Data

Quantitative implementation evaluation data can be analyzed using the following:

- Counts (e.g., 1,000 families were served over the program period)
- Averages (e.g., on average, 32 workshop sessions were provided per month)
- Frequencies (e.g., 40 percent of caseworkers had 5 or more years’ experience)

How you calculate each descriptive quantitative statistic depends on your evaluation questions. For example, you may need to calculate weekly attendance rates or monthly rates. You may need to report the mean number of workshops attended or the percentage of participants who attended at least 8 of 10 workshops.

Qualitative Data

Numerous approaches to analyzing qualitative data are available, each with different levels of rigor and requiring different levels of expertise and effort. Coding⁸ is a ubiquitous part of qualitative analysis. In general, coding processes fall into one of two categories, deductive⁹ or inductive:¹⁰

- Deductive or “theory-driven” coding is a top-down approach that applies predetermined codes.¹¹ The codes can be drawn from the literature or represent issues an evaluation team is seeking to better understand. For example, you may decide to apply the codes “transportation,” “child care,” and “work schedule” to interview transcripts with program participants based on previously reported barriers to participation. In this case, deductive coding may save time and ensure key areas of interest are coded. However, starting with predefined codes also increases the risk of researcher bias and/or could overlook other important themes.
- Inductive or “data-driven” coding is a bottom-up approach that generates codes based on the data. These codes are iteratively developed throughout a coding process that typically involves reading through the data to establish a general understanding of the issue (e.g., experience, behavior, decision, relationships), identifying meaning units,¹² assigning codes to those meaning units, and grouping codes according to themes. For example, the use of inductive coding may lead to assigning the code “cultural incongruence” to capture any participant discussion about how the program content and/or delivery may be lacking cultural sensitivity.

Both deductive and inductive strategies can be combined to facilitate a foundational understanding of the topic (from a previous evaluation of the same program, for example), while also facilitating the addition of new, unanticipated information to emerge from the data as a codebook is being developed.

Finally, qualitative analysis is not a linear process, and coding is rarely a one-time event. First-level coding mainly uses descriptive, low inference codes that are useful for summarizing segments of data (i.e., to answer questions such as who, what, when, where) and provide the basis for higher level order coding. For example, when coding an interview transcript with a program participant, any mention of gas cards or ride share payments might be coded as “supportive service payment.” Second-level codes tend to focus on patterns across multiple informants or sources of data and often require some degree of inference beyond the data.

⁸ Coding is the process of systematically categorizing excerpts in qualitative data to find themes and patterns (Delve, n.d.-a).

⁹ Deductive coding is also called a top-down approach: you start with a set of predetermined codes and then find excerpts that fit those codes (Delve, n.d.-b).

¹⁰ Inductive coding is also called a bottom-up approach: you start with no codes and develop codes as you analyze the dataset (Delve, n.d.-b).

¹¹ Codes are descriptive labels assigned to data (CESSDA Training Team, 2020).

¹² Meaning units are segments of text that describe some information about the evaluation question (Elo et al., 2014).

The previous code for “supportive service payment” might be grouped under the broader code of “participation barriers” or broken down further into subcodes such as “insufficient compensation” or “gift card challenges.”

In summary, qualitative analysis is a flexible, reflective, and continuous process of coding, recoding, and categorizing, with subsequent return to the raw data to tell a story about how program implementation occurred. Qualitative data analysis can provide insights into how planned activities occurred and why, who implemented the activities, program reach, and participant characteristics. You can then compare this information with your initial objectives and determine whether there is a difference between objectives and actual implementation. Qualitative data analysis is also used to contextualize outcome analysis findings (mixed-methods approach) as described in [chapter 5](#). This process will answer the question: Has the program been successful in attaining the anticipated implementation objectives?

If your objectives and your actual implementation differ, you can analyze your evaluation information to determine the reasons for the differences. This step answers the question: If not, why not?

You can also use your evaluation information to identify barriers that impeded implementation and facilitating factors that contributed to implementation. This information can be used to “tell the story” of your program’s implementation. Recall the measurable objectives introduced as examples in [chapter 4](#) for the planning of a substance use prevention program:

- The program will provide eight substance use education class sessions per year.
- Each session will involve 2 hours of classes per day.
- Classes will be held for 5 days.

An example of how this information might be organized is provided in table 7.2. The table represents an analysis of the program’s measurable implementation objective concerning what the program plans to do. The first column lists the measurable objectives. The actual program implementation information is provided in the second column. For this program, differences between objectives and actual implementation were apparent for three of the four measurable objectives. Column 3 notes the presence or absence of differences, and column 4 provides the reasons for those changes. Columns 5 and 6 identify the barriers encountered and the facilitating factors. These factors are important to identify regardless of whether implementation objectives were attained. They provide the context for understanding the program and will help you interpret the results of your analyses.

Table 7.2. Sample Table for Analyzing Information on Implementation Objectives

Implementation Objective	Actual Implementation	Differences? (Yes/No)	If Yes, Reasons for Change	Barriers Encountered	Facilitating Factors
Eight substance use prevention class sessions per year	Six substance use prevention class sessions the first year	Yes	Delay in startup time during the first year	Difficulty finding qualified staff Delay in curriculum development	Agency experience in implementing similar types of programs Assistance of volunteers with sessions
Each session will last 2 weeks	First two sessions lasted 2 weeks; last four sessions lasted 1 week	Yes	Participants could not consistently attend for 2 weeks	Youth lost interest during second week	Available participants in shelter
Each class will last 2 hours	First two sessions, classes were 2 hours each day; last four sessions were 3 hours each day	Yes	Because the time was shortened, had to extend intensity of classes to cover curriculum material	None	Experienced staff able to cover curriculum during shortened time span
Classes will be given 5 days of each week	5 days a week	No		Problems with crisis intervention youth attending all 5 days	Staff availability

By reviewing the information in this table, you could say the following about your program:

- The program implemented only six substance use prevention sessions instead of the intended eight sessions.
 - ▶ A delay in starting the first set of sessions caused the program to complete fewer sessions during the program evaluation timeline than expected.
 - ▶ The delay was caused by the difficulty of recruiting and hiring qualified staff, which took longer than expected.
 - ▶ With staff now on board, we expect to be able to implement the full eight sessions in the second year.
 - ▶ After staff were hired, the sessions were implemented smoothly because there were several volunteers who helped organize special events and transport participants to the events.
- For the first two sessions, the class time was 2 hours per day, as originally intended. After the number of sessions was decreased, the class time increased to 3 hours per day.
 - ▶ The increase was caused by the need to cover the curriculum material during the session.
 - ▶ The extensive experience of the staff and the assistance of volunteers facilitated covering the material during the 1-week period.
 - ▶ The youth's interest was high during the 1-week period.

- The classes were provided for 5 days, as intended.
 - ▶ This schedule was facilitated by staff availability and the access to youth residing in the shelter.
 - ▶ It was more difficult to get youth from crisis intervention services to attend for all 5 days.

You would then apply this approach to data relevant to all your other implementation objectives, such as staffing (who will do it) and the population (reach and characteristics of participants). To begin organizing the implementation information from your own program, see the blank template of table 7.2 provided in [appendix B](#).

Analyze Outcome Data

[Chapter 1](#) defines outcome evaluations as studies that intend to understand the extent to which change has occurred as intended. An impact evaluation can attribute outcomes (typically those that occur sometime after program completion) to the program. The analysis of participant outcome information typically answers the following questions:

- Did the expected changes occur in participants' knowledge, attitudes, behavior, or awareness?
- And for impact designs:
 - ▶ Did the expected changes occur in other outcomes such as participants' incomes, parenting, educational attainment, or relationship?
- If changes occurred, were they the result of your program's interventions?

If you employed a quasi-experimental or experimental impact design (and executed it well), you likely will answer questions such as the following:

- Did the program improve participant outcomes (such as increases in wages, education, or family stability, or decreases in smoking, number of missed school days, or financial hardships)?

Another question that could be included in your analysis of participant outcome information follows:

- Did some participants change more than others, and if so, what explains this difference? (For example, characteristics of the participants, types of interventions, duration of interventions, intensity of interventions, or characteristics of staff)

Your evaluation planning must include a detailed description of how you will analyze information to answer these questions. Know exactly what you want to do before you begin collecting data, particularly the types of statistical procedures you will use to analyze participant outcome information.

All outcome evaluations assess changes, so they all have a “compared with what” component. At a basic level, your analysis will calculate the value of the outcome at each postprogram time point and compare it, using a statistical procedure, with your selected comparison condition. As described in chapter 5, common comparison conditions follow:

- For nonexperimental designs: data from the same individuals before program start; benchmarks, such as national or state-level averages; or targets, such as funder expectations, or those your evaluation team develops based on evidence from other similar evaluations
- For impact (quasi-experimental and experimental) designs: outcome data on the same measures from a randomly or nonrandomly selected but similar population. To strengthen the rigor of impact studies, you can conduct difference-in-difference analyses. In difference-in-difference, you compare the change over time within your treatment group on an outcome to the change over time on the same outcome for your comparison group

Baseline equivalence

If you have an impact evaluation model, you should conduct baseline equivalence tests for each outcome at pretest data collection period. These tests are particularly important for quasi-experimental designs or random assignment designs that had high attrition. It helps to ensure that treatment and comparison groups were equal on outcomes of concern before the treatment group received the program.

Understanding statistical procedures

For outcome analyses, you will conduct inferential analyses. Unlike descriptive analyses (mentioned above) that aim to describe, inferential analyses aim to test relationships among data elements.

Inferential analyses

- **Measure the degree to which the outcome variable and other variables are associated.** At a basic level, you will test the relationship between the outcome of interest and one or more independent variables¹³ (such as characteristics of the program). This type of analysis can indicate whether the levels of individuals' outcomes are correlated with the independent variables. For example, you may determine that outcomes are positively correlated with hours of program services received. Remember that correlation does not mean causation. Researchers can include additional independent variables to “control” the analysis for factors that are associated with both the outcome and the program characteristic. Factors such as socioeconomic and demographic characteristics and “pretest” measures of the outcome often serve as strong control variables. Impact analyses will contrast individuals offered the program with individuals not offered the program by including a program indicator as an independent variable.

¹³ An independent variable is one that stands alone and isn't changed by the other variables being measured (National Center for Education Statistics, n.d.)

- **Can be conducted using multiple types of models.** It is important the evaluation expert on your team knows which types of tests are appropriate for which types of data. For example, ordinary least squares regression models work with continuous (e.g., weight) or ordinal (e.g., 5-point attitude scale) dependent variables, while logistic regressions are typically used to test binary dependent variables (e.g., did or did not drop out of school). In contrast, an ANOVA test is applicable for categorical dependent variables (e.g., when the outcome of interest is a set of different categories that can't be ranked, such as marital status). Your expert should confirm your data meets the assumptions of proposed models, such as normal distribution for a multiple regression model.
- Produce two important pieces of information:
 - ▶ Statistical significance: The p -value is the probability that an impact greater than your estimate would occur by chance when the true impact is zero. See U.S. Department of Education (2021) for a more technical explanation.
 - ▶ Measure of magnitude: an indication of how much change in your outcome (dependent variable) occurred. Impact analyses often use a calculation called an effect size. You can also demonstrate magnitude by calculating difference in outcomes through the model (e.g., on average, program participants scored 10 points higher on a parenting measure at program completion compared with pretest). You can also report on clinical or meaningful changes; for example, of the children in the program with below grade-level reading skills before enrollment, half scored at or above grade level at posttest.

After you have answered your primary evaluation questions, you may want to conduct some exploratory analyses, such as the following:

- **Subgroup analyses.** In subgroup analyses, you test to see if certain participants experienced larger or smaller differences in an outcome. For example, you may have a hunch, or a hypothesis, that women who participate in your parenting program were more likely to report increased confidence in their parenting than men. A subgroup analysis would separate the outcome change for those two groups and help verify or refute your hunch.
- **Dosage analyses.** In dosage analyses, you can assess the extent to which individuals who received more of your program had better outcomes than those who received less. Be careful with the interpretation of these findings. People who take more programming may differ in important ways from those who take less programming, and those differences might drive changes in outcomes as opposed to the program. For example, these two groups could differ on motivation to change, access to transportation, health, or level of stress.
- **Sensitivity analyses.** Sensitivity analyses are approaches that test the robustness of your model (and its findings). In sensitivity analyses, you change the assumption in your model to see the extent to which slightly different assumptions lead to different findings.

Interpreting Your Findings

The interpretation step of program evaluation might be the most meaningful. When you make sense of the findings, you can tell the story of your program through the evaluation results. Without bringing meaning to these numbers and tables, you can't make use of the evaluation. With evaluation, you can explain how your program will perform to future funders, make program adaptations and improvements, provide potential program participants with evidence about what their experience might be like, and potentially scale and replicate your program in other locations with fidelity.

Interpretation involves setting down program information and evaluation results and asking questions such as the following:

- Was this finding what we expected? Why or why not?
- Why did one outcome show statistical improvement? Why didn't others?
- Do we think the program is responsible for this outcome change? What else could be affecting change at the same time?
- What did we do well? What do we need to change or do better?

Information to have at hand during interpretation efforts

During interpretation conversations and efforts, you should be able to reference, compare, and revisit much relevant program information. To be fully prepared for that conversation, you should gather important materials such as your program logic model, staff training materials, funder requirements, program objectives and goals, and any other relevant information such as evaluations of similar programs.

Findings from both your outcome and implementation evaluation should come together during this process. For example, if you did not see expected improvements in participant knowledge of community resources for social support, you should look toward your implementation evaluation findings. Your implementation evaluation found that most program facilitators were new to the community and didn't have a strong understanding of the different community-based organizations in the area.

Pay attention to the magnitude of your findings. While general evaluation practice suggests evaluation teams should focus on findings that demonstrate statistical significance, discussing the magnitude of those statistically significant findings brings important nuance to your evaluation. Large, statistically significant findings are more meaningful with respect to real-life improvement than small ones. At the same time, there's an adage, "The absence of evidence is not necessarily the evidence of absence." In other words, spend time exploring why you might have received unexpected, nonsignificant findings. If you can't tie that finding to implementation problems, it could be a function of your evaluation design or evaluation quality. Reassess whether you had a large enough sample size (it's easier to find statistical significance with larger samples), used an appropriate measure, measured the outcome at the right time to see change, or had a low response rate (which may have skewed which types of people provided outcome data).

Throughout the interpretation process, keep methodological and situational limitations in mind. For example, a large magnitude of change measured by a survey with low response rates should be viewed with great

caution because the underlying quality of the data could be problematic. Your evaluation design affects the extent to which outcome changes can be seen as caused by the program. In addition, situate your findings within the current evaluation context. Many of your evaluation audience members will want to know if your findings are likely to happen in other situations. For example, if your program showed significant improvements in reading scores, they may want to know if those improvements would occur if they implemented the same program. Be sure to think about and document the context in which your program operated, and the extent to which you think certain elements were key to your findings.

Practice Culturally Responsive and Equitable Evaluation When Analyzing and Interpreting Data

A CREE approach does not stop with the evaluation design or data collection, but it is also helpful when analyzing and interpreting the data. Numerous ways are available to meaningfully engage community members during the analysis stage to ensure a more robust understanding of the data. For example, data walks, a strategy for visually sharing data with community members, create an opportunity for program participants, community residents, and service providers to jointly review data presentations in small groups, interpret what the data mean, and collaborate to use their individual expertise to improve policies, programs and other factors of community change (Murray et al., 2015).

Understanding when and how to disaggregate data is also a valuable way to practice CREE during data analysis. This practice enables evaluators to identify and address findings for groups of participants, instead of just participants as a whole. Data should be disaggregated to the level where it can still be understood in a meaningful way for the group it is exploring. For example, can you look at data by race, age, and gender to understand if there is a commonality in experience for Asian American females over 60? Data disaggregation can be informative when reviewing or analyzing data, conducting root cause analyses, reporting findings, or presenting information. Before analyzing data at a subgroup level, think about what types of bias could reside in the measurement tool or data collection process that could influence differences between groups. Carefully frame your findings so you don't inadvertently reinforce racial stereotypes.

A CREE approach is vital when interpreting the results of your analyses. Evaluation teams should consider engaging program participants and community members when interpreting analyses to ensure conclusions drawn are informed by the community's cultural values and perspectives of the program's quality and effectiveness. Evaluation teams should also be thoughtful about contextualizing research evidence to the local settings and systems (e.g., examining historical or structural factors that might shape findings). This can be particularly important when interpreting negative or unexpected findings. More specifically, evaluators must be careful to draw conclusions that account for both individual and system-level factors that could contribute to negative findings.

In summary, consider strategies to promote CREE practices both when analyzing and interpreting the data they've collected and the following recommendations:

- Disaggregate data and findings to illuminate disproportionality.
- Contextualize data using information about lived experiences.
- Include findings that communicate assets and strengths within a community.
- Provide context for findings, such as relevant institutional and environmental factors that influence individual behaviors.
- Ensure community voices are heard when making judgments about the data.
- Demonstrate cultural humility when interpreting findings.

Analyzing and interpreting data through a culturally responsive and equitable lens

Consider the first-time parent-child development education program discussed earlier in this chapter. Disaggregating the data may reveal your child development program is effective for White parents but not for parents of color. In such a case, you could examine your analysis of program implementation information to understand why this may have happened and provide recommendations for program improvement. By contrast, if you had not disaggregated the data in this manner, this disproportionality likely would not have been unearthed and addressed.

If you find the program is effective for White parents but not for parents of color, have you considered the potential role of structural racism? For instance, structural racism may have a negative effect on the number of instructors of color, which in turn, may reduce the program's effectiveness for parents of color.

To learn more ...

- [Analyzing and Interpreting Data](#) (Wilder Research, 2009)
- [Analyzing Quantitative Data for Evaluation](#) (CDC, 2018)
- [Disaggregating Data by Race Allows for More Accurate Research](#) (Sharpe, 2019)
- [Ethics and Empathy in Using Imputation to Disaggregate Data for Racial Equity: Recommendations and Standards Guide](#) (Brown et al., 2021)
- [Forum Guide to Collecting and Using Disaggregated Data on Racial/Ethnic Subgroups](#) (National Forum on Education Statistics, n.d.)
- [Methods, Challenges, and Best Practices for Conducting Subgroup Analysis](#) (Breck & Wakar, 2021)
- [Practical Strategies for Culturally Competent Evaluation](#) (CDC, 2014)
- [Qualitative Methods in Monitoring and Evaluation: Analyzing Qualitative Data](#) (Peters, 2022)
- [The Essentials of Disaggregated Data for Advancing Racial Equity](#) (Race Matters Institute, 2019)

References

- ACF (Administration for Children and Families). (n.d.-a). *Title IV-E prevention services clearinghouse*. U.S. Department of Health and Human Services. <https://preventionservices.acf.hhs.gov/>
- ACF. (n.d.-b). *Home visiting evidence of effectiveness*. U.S. Department of Health and Human Services. <https://homvee.acf.hhs.gov/>
- Asher, J., Resnick, D., Brite, J., Brackbill, R., & Cone, J. (2020). An introduction to probabilistic record linkage with a focus on linkage processing for WTC registries. *International Journal of Environmental Research and Public Health*. <https://doi.org/10.3390/ijerph17186937>
- Bailey, J. (2008). First steps in qualitative data analysis: Transcribing. *Family Practice*, 25(2), 127–131. <https://academic.oup.com/fampra/article/25/2/127/497632>
- Breck, A., & Wakar, B. (2021). *Methods, challenges, and best practices for conducting subgroup analysis* (OPRE Report 2021-17). 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/methods-challenges-and-best-practices-conducting-subgroup-analysis>
- Brown, S., Ford, L. D., & Ashley, S. (2021). *Ethics and empathy in using imputation to disaggregate data for racial equity: recommendations and standards guide*. Urban Institute. <https://www.urban.org/research/publication/ethics-and-empathy-using-imputation-disaggregate-data-racial-equity-recommendations-and-standards-guide>
- CDC (Centers for Disease Control and Prevention). (2014). *Practical strategies for culturally competent evaluation*. U.S. Department of Health and Human Services. https://www.cdc.gov/asthma/program_eval/cultural_competence_guide.pdf
- CDC. (2018). *Analyzing quantitative data for evaluation*. *Evaluation Brief*. U.S. Department of Health and Human Services. <https://www.cdc.gov/healthyyouth/evaluation/pdf/brief20.pdf>
- CESSDA Training Team. (2017–2020). *Qualitative coding*. <https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide/3.-Process/Qualitative-coding>
- Deke, J., Sama-Miller, E., & Hershey, A. (2015). *Addressing attrition bias in randomized controlled trials: Considerations for systematic evidence reviews*. U.S. Department of Health and Human Services, Administration for Children and Families, Office of Planning, Research, and Evaluation. https://homvee.acf.hhs.gov/sites/default/files/2019-06/HomVEE-Attrition-White_Paper-7-2015.pdf
- Delve. (n.d.-a). *The essential guide to coding qualitative data*. <https://delvetool.com/guide#:~:text=Qualitative%20coding%20is%20a%20process,themes%20and%20patterns%20for%20analysis>.
- Delve. (n.d.-b). *Deductive and inductive coding*. <https://delvetool.com/blog/deductiveinductive#:~:text=Inductive%20coding%20is%20a%20ground,from%20the%20raw%20data%20itself>

- Elo, S., Kääriäinen, M., Kanste, O., Pölkki, T., Utriainen, K., & Kyngäs, H. (2014). *Qualitative content analysis: A focus on trustworthiness*. SAGE Open. <https://doi.org/10.1177/2158244014522633>
- Multon, K. D., & Coleman, J. S. M. (2018). Inter-rater reliability. In B. B. Frey (Ed.), *The SAGE encyclopedia of educational research, measurement, and evaluation*. <https://methods.sagepub.com/reference/the-sage-encyclopedia-of-educational-research-measurement-and-evaluation/i11331.xml>
- Murray, B., Falkenburger, E., & Saxena, P. (2015). *Data walks: An innovative way to share data with communities*. Urban Institute. <https://www.urban.org/research/publication/data-walks-innovative-way-share-data-communities>
- National Center for Education Statistics. (n.d.). *What are independent and dependent variables?* [graphing tutorial]. https://nces.ed.gov/nceskids/help/user_guide/graph/variables.asp
- National Forum on Education Statistics. (n.d.). *Forum guide to collecting and using disaggregated data on racial/ethnic subgroups*. https://nces.ed.gov/forum/pdf/Disaggregated_Data_PPT.pdf
- OPRE (Office of Planning, Research, and Evaluation). (2010). *The program manager's guide to evaluation* (Second ed.). 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>
- Peters, B. (2022). *Qualitative methods in monitoring and evaluation: Analyzing qualitative data*. <https://programs.online.american.edu/msme/masters-in-measurement-and-evaluation/resources/qualitative-methods-project-cycle>
- Race Matters Institute of JustPartners. (2019). *The essentials of disaggregated data for advancing racial equity*. <https://viablefuturescenter.org/racemattersinstitute/resources/disaggregated-data/>
- Robert Wood Johnson Foundation. (2006). *Qualitative research guidelines project*. <http://www.qualres.org/HomeAudi-3700.html#:~:text=An%20audit%20trail%20is%20a,was%20done%20in%20an%20investigation>
- Satterlund, T. (n.d.). *Note to self: Writing analytic memos*. UC Davis Center for Evaluation and Research. https://tobaccoeval.ucdavis.edu/sites/g/files/dgvnsk5301/files/inline-files/Newsletter--Memoing%202.22.12_edited.pdf
- Sharpe, R. (2019). Disaggregating data by race allows for more accurate research. *Nature Human Behaviour*, 3, 1240. <https://doi.org/10.1038/s41562-019-0696-1>
- Statistics Solutions. (n.d.). *Thematic analysis*. <https://www.statisticssolutions.com/thematic-analysis/#:~:text=Familiarization%3A%20This%20is%20the%20process,qualitative%20researcher%20with%20the%20data>
- UCLA. (n.d.). *What does Cronbach's alpha mean?* SPSS FAQ. Advanced Research Computing. <https://stats.oarc.ucla.edu/spss/faq/what-does-cronbachs-alpha-mean/>

University of Melbourne. (n.d.). *Reflexivity*. <https://medicine.unimelb.edu.au/school-structure/medical-education/research/qualitative-journey/themes/reflexivity#:~:text=Reflexivity%20is%20about%20acknowledging%20your,will%20influence%20the%20research%20process>

U.S. Department of Education. (n.d.) *What works clearinghouse*. <https://ies.ed.gov/ncee/wwc/>

U.S. Department of Education. (2021). *Statistical significance and sample size*. National Center for Educational Statistics. <https://nces.ed.gov/nationsreportcard/guides/statsig.aspx>

U.S. Department of Labor. (n.d.). CLEAR: *Clearinghouse for labor evaluation and research*. <https://clear.dol.gov/#:~:text=Homepage%20%7C%20CLEAR&text=CLEAR's%20mission%20is%20to%20make,about%20labor%20policies%20and%20programs>

Wilder Research. (2009). *Analyzing and interpreting data: Evaluation resources from Wilder Research*. <http://www.evaluatod.org/assets/resources/evaluation-guides/analyzing-interpretingdata-8->