IES Institute of Education Sciences

Program Evaluation Toolkit: Quick Start Guide

Regional Educational Laboratory Central

At Marzano Research

REL 2022–112
U.S. DEPARTMENT OF EDUCATION

A Publication of the National Center for Education Evaluation and Regional Assistance at IES

Program Evaluation Toolkit: Quick Start Guide

Joshua Stewart, Jeanette Joyce, Mckenzie Haines, David Yanoski,
Douglas Gagnon, Kyle Luke, Christopher Rhoads, and Carrie Germeroth

October 2021

Program evaluation is important for assessing the implementation and outcomes of local, state, and federal programs. The Program Evaluation Toolkit provides tools and resources to support individuals responsible for evaluating and monitoring local, state, or federal programs. The toolkit comprises eight modules that cover critical steps in program evaluation, beginning at the planning stages and progressing to the presentation of findings.

Contents

Unpacking the Program Evaluation Toolkit

What is the toolkit?

The Program Evaluation Toolkit presents a step-by-step process for conducting a program evaluation. Program evaluation is important for assessing the implementation and outcomes of local, state, and federal programs. Designed to be used in a variety of education settings, the toolkit focuses on the practical application of program evaluation. The toolkit can also build your understanding of program evaluation so that you can be better equipped to understand the evaluation process and use evaluation practices.

The toolkit consists of this *Quick Start Guide* and a website with eight modules that begin at the planning stages of an evaluation and progress to the presentation of findings to stakeholders. Each module covers a critical step in the evaluation process.

The toolkit is available at https://ies.ed.gov/ncee/edlabs/regions/central/resources/pemtoolkit/index.asp.

The toolkit includes a screencast that provides an overview of each stage of the evaluation process. It also includes tools, handouts, worksheets, and a glossary of terms (see the appendix of this guide) to help you conduct your own evaluation. The toolkit resources will help you create a logic model, develop evaluation questions, identify data sources, develop data collection instruments, conduct basic analyses, and disseminate findings.

What is program evaluation?

Program evaluation is the systematic process for planning, documenting, and assessing the implementation and outcomes of a program. Evaluations often address the following questions:

- Is the program effective?
- Can the program be improved?

A well-thought-out evaluation can identify barriers to program effectiveness, as well as catalysts for program successes. Program evaluation begins with outlining the framework for the program, determining questions about program milestones and goals, identifying what data address the questions, and choosing the appropriate analytical method to address the questions. By the end, an evaluation should provide easy-to-understand findings, as well as recommendations or possible actions.

Who should use the toolkit?

The primary audience for the toolkit is individuals who evaluate local, state, or federal programs. Other individuals engaged in program evaluation might also benefit from the toolkit. The toolkit will be particularly helpful to individuals responsible for:

- Designing evaluations of program implementation and outcomes.
- Collecting and analyzing data about program implementation and outcomes.
- Writing reports or disseminating information about program implementation and outcomes.

Am I ready to use this toolkit?

This toolkit covers the main components of program evaluation, from foundational practices to quantitative and qualitative methods, to dissemination of findings. The toolkit content is broad and might challenge you to think in new ways. However, you do not need prior experience or advanced training in program evaluation to benefit from using the toolkit. In addition to the main content for general users, optional resources in the toolkit can help more advanced users refine their knowledge, skills, and abilities in program evaluation.

The following questions can help you determine your readiness to use the toolkit without support from colleagues:

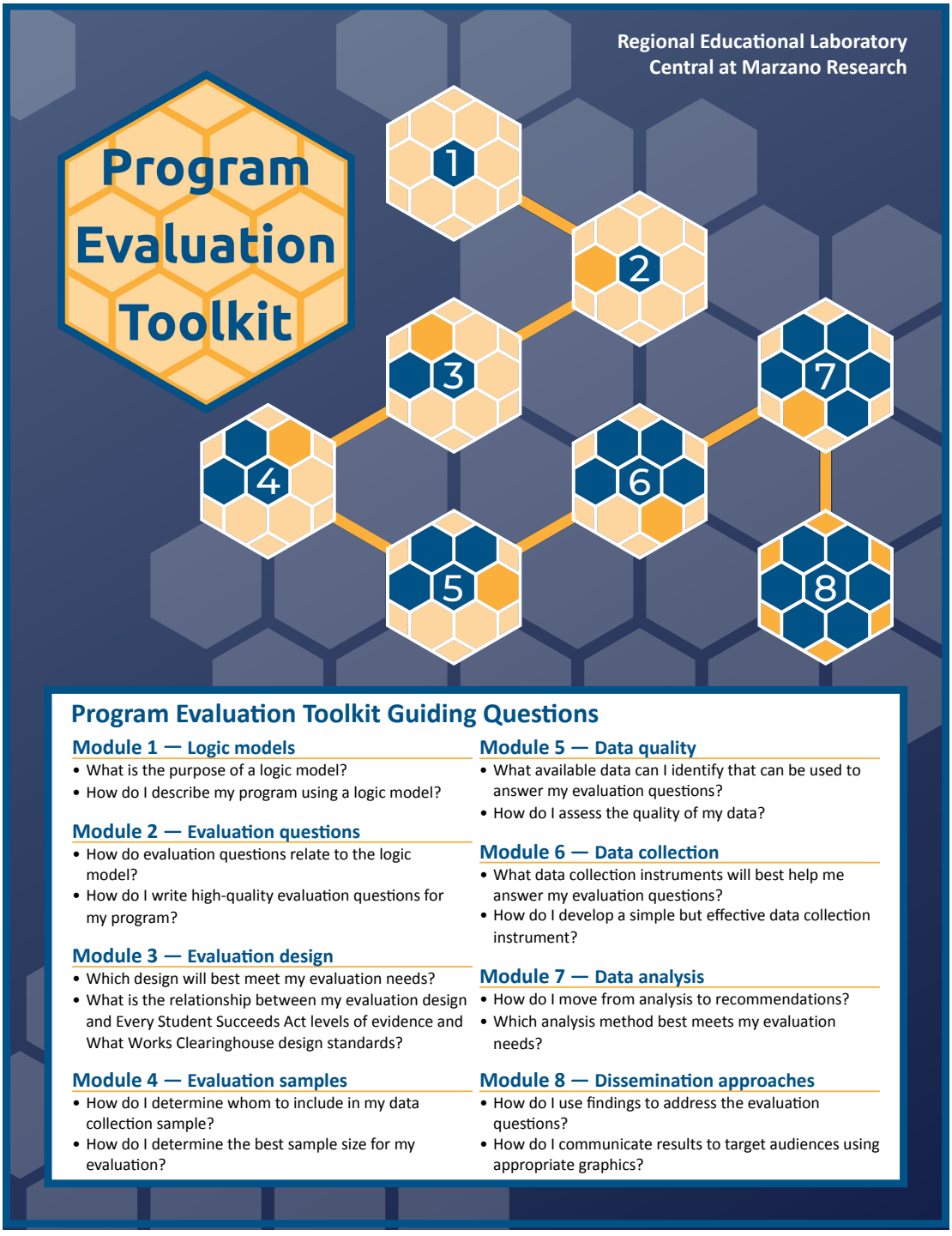
- Are you, or have you been, engaged in program evaluation?
- Do you have basic data literacy, gained from some experience in gathering, reviewing, and using data?

Where do I start?

You can progress through the toolkit modules either sequentially or selectively, reviewing only modules that pertain directly to your current evaluation needs (figure 1). In each module the first chapter provides a basic introduction to the module topic, and the subsequent chapters increase in complexity and build on the basic introduction. For each module you can decide what level of complexity best meets your program evaluation needs. Modules, 3, 4, and 7 require statistical knowledge. If you lack statistical expertise, you might consider working through them with a colleague who has statistical expertise. You can use the toolkit tracker to document your progress (figure 2). In the tracker you can record when you start a module and which modules you have completed.

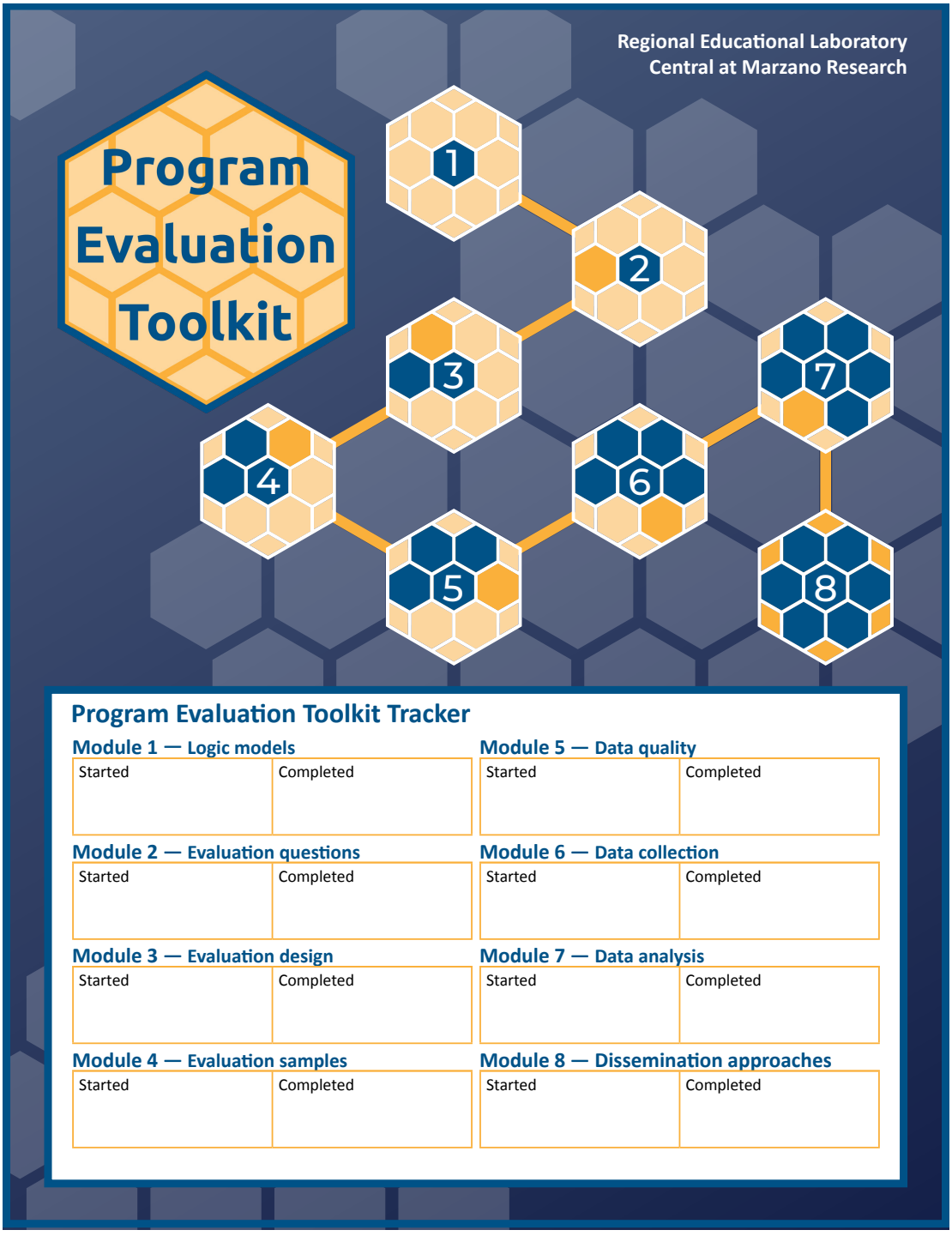
It is best to start with Module 1: Logic models, which focuses on developing a logic model for your program. A logic model is a graphical representation of the relationship between program components and desired outcomes. A well-crafted logic model will serve as the foundation for the other modules in the toolkit. You will draw on your logic model when developing measurable evaluation questions, identifying quality data sources, and selecting appropriate analyses and other key components of your evaluation. If you choose to progress through the toolkit selectively, the module selection checklist can help you identify which modules to prioritize (table 1).

Figure 1. Guiding questions for the Program Evaluation Toolkit



Source: Authors' creation.

Figure 2. Tracker for the Program Evaluation Toolkit



Source: Authors' creation.

Table 1. Module selection checklist

Module	What are my evaluation needs?	
1. Logic models	I need to clearly define my program and my expected outcomes.	☐
2. Evaluation questions	I need to develop or refine a set of relevant and measurable evaluation questions.	☐
3. Evaluation design	I need to identify an evaluation design that will ensure that claims made from my evaluation are justifiable and align to tiers of evidence under the Every Student Succeeds Act and What Works Clearinghouse design standards.[a]	☐
4. Evaluation samples	I need to determine which participants (for example, students, parents) and how many to include in my evaluation.[a]	☐
5. Data quality	I need to identify available data (for example, state assessments, student attendance) to address my evaluation questions and assess the quality of the available data.	☐
6. Data collection	I need to develop or identify quality instruments (for example, focus group protocols or surveys) to collect additional data.	☐
7. Data analysis	I need to analyze my data and make recommendations for next steps to decisionmakers.[a]	☐
8. Dissemination approaches	I need to share the findings of my evaluation with different audiences (for example, teachers, community members).	☐

a. This module includes technical information and might require more advanced statistical knowledge.

Source: Authors' compilation.

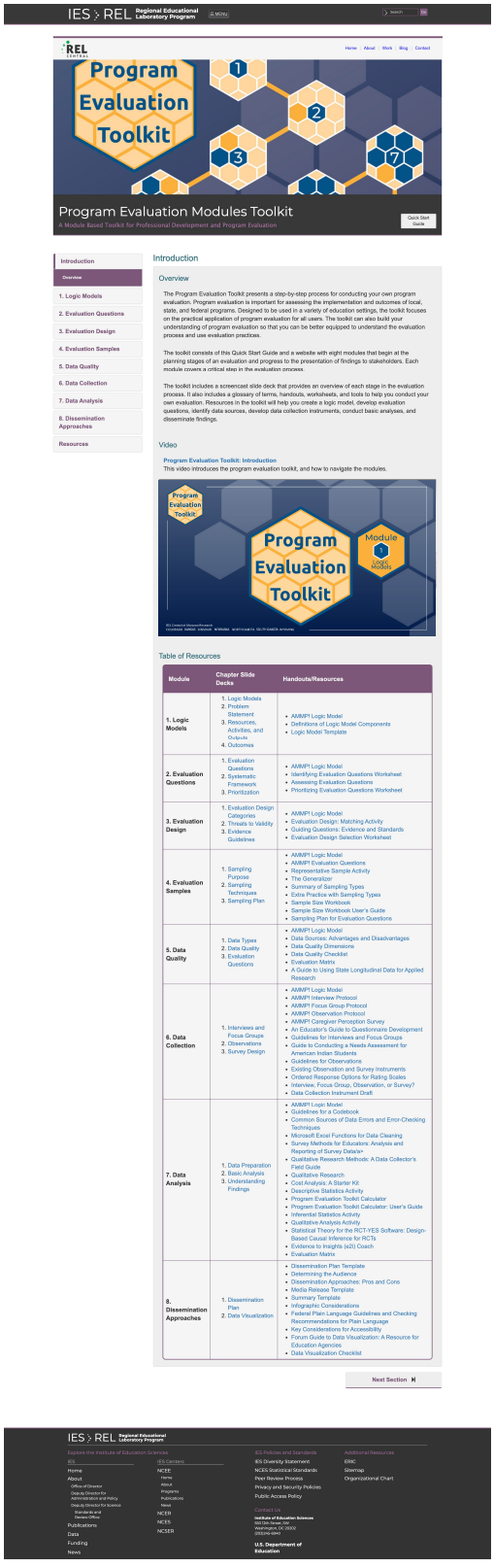
What is not included in the toolkit?

This toolkit provides tools and resources for general program evaluation. Although the toolkit can help you establish a common language around program evaluation and use resources for basic evaluation purposes, it does not include detailed information on topics, such as the advanced statistical methods of regression discontinuity designs, difference in differences designs, propensity score matching, crossover designs, and multilevel modeling. Instead, the toolkit will support you in executing simpler designs and analyses using widely available software and materials. The toolkit is designed for individuals with a basic understanding of data, statistics, and evaluation. If your evaluation requires more complex methodologies or analyses, consider consulting an evaluation expert at a university or college, reaching out to the Regional Educational Laboratory in your region, or checking out additional resources, such as the free software RCT-YES.

How do I navigate the toolkit website?

When you first open the Program Evaluation Toolkit website, you will find an introduction to the toolkit and links to each of the eight modules (figure 3).

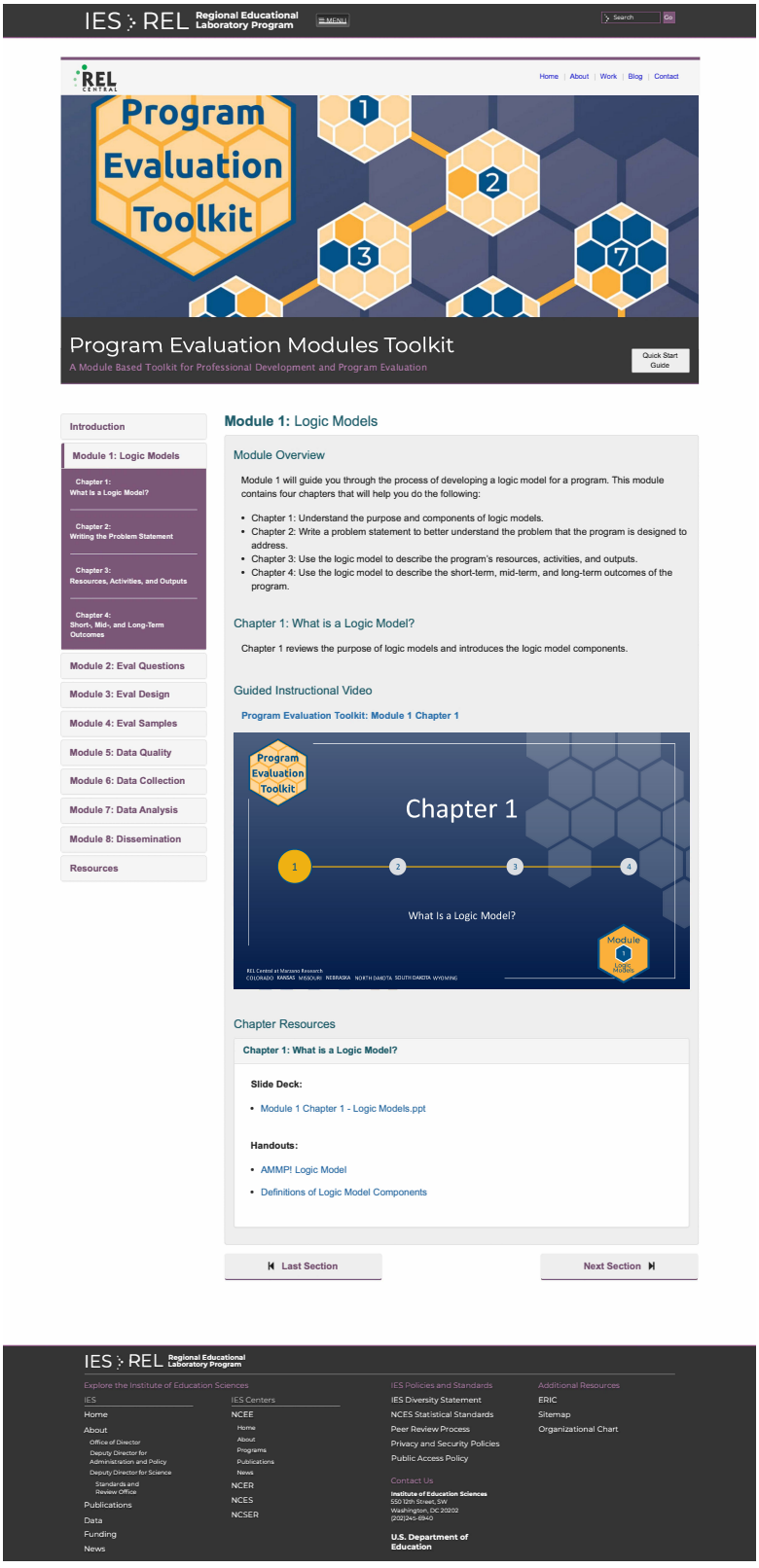
Figure 3. Opening page of the Program Evaluation Toolkit website



Source: Authors' creation.

Clicking on any of the eight module links will bring you to a webpage with information about the module content, organized into chapters (figure 4). You can use the chapters to engage with the module content in smaller sections. Each chapter includes a short video that explains the content and a link to the PowerPoint slides used in the video. In addition, each module webpage includes links to the tools, handouts, and worksheets used in the module. You can download and print these materials to use while watching the video, or you can use them while conducting your own evaluation.

Figure 4. Opening page of Module 1 on the Program Evaluation Toolkit website



Source: Authors' creation.

What is included in the toolkit?

The following sections provide short overviews of the eight modules in the toolkit. For clarification, key terms are linked to their glossary definitions in the appendix of this guide.

Module 1: Logic models

Viewing time: 36 minutes

Module 1 guides you through developing a logic model for a program. The module contains four chapters that will help you do the following:

- Chapter 1: Understand the purpose and components of logic models.
- Chapter 2: Write a problem statement to better understand the problem that the program is designed to address.
- Chapter 3: Use the logic model to describe the program's resources, activities, and outputs.
- Chapter 4: Use the logic model to describe the short-term, mid-term, and long-term outcomes of the program.

Chapter 1 reviews the purpose of logic models and introduces the logic model components. Chapter 2 explains how to write a problem statement that describes the reason and context for implementing the program. Chapters 3 and 4 present the central logic model components: resources, activities, outputs, and short-term, mid-term, and long-term outcomes. These two chapters also explain how the components relate to and inform the overall logic model. In addition, the module highlights available resources on logic model development.

Module 2: Evaluation questions

Viewing time: 37 minutes

Module 2 guides you through writing measurable evaluation questions that are aligned to your logic model. The module contains three chapters that will help you do the following:

- Chapter 1: Learn the difference between process and outcome evaluation questions and understand how they relate to your logic model.
- Chapter 2: Use a systematic framework to write, review, and modify evaluation questions.
- Chapter 3: Prioritize questions to address in the evaluation.

Chapter 1 introduces the two main types of evaluation questions (process and outcome) and explains how each type aligns to the logic model. Chapter 2 presents a systematic framework for developing and revising evaluation questions and then applies that framework to sample evaluation questions. Chapter 3 describes and models a process for prioritizing evaluation questions. The module includes worksheets to help you write, review, and prioritize evaluation questions for your own program.

Module 3: Evaluation design

Viewing time: 37 minutes

Module 3 reviews major considerations for designing an evaluation. The module contains three chapters that will help you understand the following:

- Chapter 1: The major categories of evaluation design, including when to use each design.
- Chapter 2: Threats to validity, including how to consider these threats when designing an evaluation.
- Chapter 3: The relationship between evaluation design and Every Student Succeeds Act (ESSA) tiers of evidence and What Works Clearinghouse (WWC) design standards.

Chapter 1 introduces four major categories of evaluation design: descriptive designs, correlational designs, quasi-experimental designs, and randomized controlled trials. The chapter explains considerations for when to use each category, including which is suited to the two types of evaluation questions (see module 2). Chapter 2 presents threats to internal and external validity and provides examples of common challenges in designing evaluations. Chapter 3 discusses the four tiers of evidence in ESSA and the three ratings of WWC design standards. The chapter explains how each tier or rating connects to evaluation design choices. The module includes activities to help you identify appropriate evaluation designs and links to resources from which you can learn more about the ESSA tiers of evidence and WWC design standards.

Module 4: Evaluation samples

Viewing time: 57 minutes

Module 4 provides an overview of sampling considerations in evaluation design and data collection. The module contains three chapters that will help you understand the following:

- Chapter 1: The purpose and importance of sampling.
- Chapter 2: Sampling techniques that you can use to obtain a desirable sample.
- Chapter 3: Methods for determining sample size and for creating a sampling plan for your evaluation.

Chapter 1 reviews the purpose of sampling and defines key terms, including representativeness, generalizability, and weighting. The chapter also details the process for selecting a representative sample. Chapter 2 covers the different types of random and nonrandom sampling techniques. Chapter 3 introduces a tool for determining the optimal sample size, as well as a process for drafting a sampling plan.

Module 5: Data quality

Viewing time: 30 minutes

Module 5 provides an overview of data quality considerations. The module also covers aligning data to evaluation questions. The module contains three chapters that will help you do the following:

- Chapter 1: Identify the two major types of data and describe how to use them in an evaluation.
- Chapter 2: Evaluate the quality of your data, using six key criteria.
- Chapter 3: Connect data to your evaluation questions.

Chapter 1 discusses the two main types of data (quantitative and qualitative) and explains how to use both types of data to form a more complete picture of the implementation and outcomes of your program. Chapter 2 discusses the key elements of data quality: validity, reliability, timeliness, comprehensiveness, trustworthiness, and completeness. In addition, the chapter includes a checklist for assessing the quality of data. Chapter 3 covers the alignment of data to evaluation questions. The chapter introduces the evaluation matrix, a useful tool for planning your evaluation and the data you need to collect.

Module 6: Data collection

Viewing time: 42 minutes

Module 6 presents best practices in developing data collection instruments and describes how to create quality instruments to meet data collection needs. The module contains three chapters that will help you do the following:

- Chapter 1: Plan and conduct interviews and focus groups.
- Chapter 2: Plan and conduct observations.
- Chapter 3: Design surveys.

Chapter 1 describes how to prepare for and conduct interviews and focus groups to collect data to answer evaluation questions. Chapter 2 covers developing and using observation protocols that include, for example, recording checklists and open field notes, to collect data. Chapter 3 focuses on survey development and implementation. Each chapter includes guiding documents, examples of data collection instruments, and a step-by-step process for choosing and developing an instrument that best meets your evaluation needs.

Module 7: Data analysis

Viewing time: 53 minutes

Module 7 reviews major considerations for analyzing data and making recommendations based on findings from the analysis. The module contains three chapters that will help you understand the following:

- Chapter 1: Common approaches to data preparation and analysis.
- Chapter 2: Basic analyses to build analytic capacity.
- Chapter 3: Implications of findings and how to make justifiable recommendations.

Chapter 1 reviews common techniques for data preparation, such as identifying data errors and cleaning data. It then introduces quantitative methods, including basic descriptive methods and linear regression. The chapter also reviews basic qualitative methods. Chapter 2 focuses on cleaning and analyzing quantitative and qualitative datasets, applying the methods from chapter 1. Chapter 3 presents a framework and guiding questions for moving from analysis to interpretation of the findings and then to making defensible recommendations based on the findings.

Module 8: Dissemination approaches

Viewing time: 47 minutes

Module 8 presents best practices in disseminating and sharing the evaluation findings. The module contains two chapters that will help you do the following:

- Chapter 1: Learn how to develop a dissemination plan.
- Chapter 2: Explore best practices in data visualization.

Chapter 1 describes a dissemination plan and explains why a plan is helpful for sharing evaluation findings. It then outlines key considerations for developing a dissemination plan, such as the audience, the message, the best approach for communicating the message, and the best time to share the information with the audience. The chapter also includes important considerations for ensuring that dissemination products are accessible to all members of the audience. Chapter 2 reviews key considerations for visualizing data, including the audience, message, and approach. The chapter also presents examples of data visualizations, including graphs, charts, and tables, that can help make the data more easily understandable.

How did stakeholders collaborate in developing the toolkit?

The development of this toolkit arose in response to the Colorado Department of Education's need for tools and procedures to help districts systematically plan and conduct program evaluations related to locally implemented initiatives. The Regional Educational Laboratory Central partnered with the Colorado Department of Education to develop an evaluation framework and a set of curated resources that cover program evaluation from the planning stages to presentation of findings. The Program Evaluation Toolkit is an expansion of this collaborative work.

Appendix. Glossary of terms

This appendix provides definitions of key terms used in the Program Evaluation Toolkit. Terms are organized by module and listed in the order in which they are introduced in each module.

Module 1: Logic models

Logic model: A graphical representation of the relationship between the parts of a program and its expected outcomes.

Problem statement: A description of the problem that the program is designed to address.

Resources: All the available means to address the problem, including investments, materials, and personnel. Resources can include human resources, monetary resources, facilities, expertise, curricula and materials, time, and any other contributions to implementing the program.

Activities: Actions taken to implement the program or address the problem. Activities can include professional development sessions, after-school programs, policy or procedure changes, use of a curriculum or teaching practice, mentoring or coaching, and development of new materials.

Outputs: Evidence of program implementation. Outputs can include required deliverables, the number of activities, newly developed materials, new policies or procedures, observations of the program in use, numbers of students or teachers involved, and other data that provide evidence of the implementation of activities in the program.

Outcomes: The anticipated results once you implement the program. Outcomes are divided into three types:

> **Short-term outcomes:** The most immediate results for participants that can be attributed to program activities. Short-term outcomes are typically changes in knowledge or skills. Short-term outcomes are expected immediately following exposure to the program (or shortly thereafter).
>
> **Mid-term outcomes:** The more distant, though anticipated, results of participation in program activities that require more time to achieve. Mid-term outcomes are typically changes in attitudes, behaviors, and practices. Mid-term results are expected after the participants in the program have had sufficient time to implement the knowledge and skills that they have learned.
>
> **Long-term outcomes:** The ultimately desired outcomes from implementing program activities. Long-term results are expected after the changes in attitudes, behaviors, and practices have been in place for a sufficient period of time. They are typically

systemic changes or changes in student outcomes. They might not be the sole result of the program, but they are associated with it and might manifest themselves after the program concludes.

Additional considerations: Important details or ideas that do not fit into the other components of the logic model. Additional considerations can include assumptions about the program, external factors not covered in the problem statement, and factors that might influence program implementation but are beyond the evaluation team's control.

Module 2: Evaluation questions

Evaluation questions: The questions that the evaluation is designed to answer. Evaluation questions typically focus on promoting program improvement or determining the impact of a program. There are two main types of evaluation questions:

> **Process questions:** Questions about the quality of program implementation and improvement. They are also called formative questions.
>
> **Outcome questions:** Questions about the impact of a program over time. They are also called summative questions.

PARSEC: A framework for creating quality evaluation questions. PARSEC is an acronym for pertinent, answerable, reasonable, specific, evaluative, and complete.

> **Pertinent:** A question is strongly related to the information that program stakeholders and participants want to obtain from an evaluation. Pertinent questions are derived from the logic model.
>
> **Answerable:** The data needed to answer a question are available or attainable.
>
> **Reasonable:** A question is linked to what a program can practically and realistically achieve or influence.
>
> **Specific:** A question directly addresses a single component of the logic model. Specific questions are clearly worded and avoid broad generalizations.
>
> **Evaluative:** The answer to a question is actionable. Evaluative questions can inform changes to a program, policy, or initiative.
>
> **Complete:** The entire set of questions addresses all the logic model components that are of critical interest.

Important/urgent rating system: A strategy for prioritizing evaluation questions based on their importance and urgency.

> **Importance:** An important question is necessary to improve or assess a program.
>
> **Urgency:** An urgent question needs an answer as soon as possible, either to satisfy reporting requirements or to obtain necessary information before moving forward.

Module 3: Evaluation design

Evaluation design: The data collection processes and analytic methods used to answer the evaluation questions. An evaluation design should be informed by the program goals, logic model, evaluation questions, available resources, and funding requirements. There are four broad categories of evaluation design:

Descriptive designs: Used to describe a program by addressing "who," "what," "where," "when," and "to what extent" questions as they relate to the program.

Correlational designs: Used to identify a relationship between two variables and determine whether that relationship is statistically meaningful. Correlational analyses do not demonstrate causality. They can find that *X* is related to *Y,* but they cannot find that *X* caused *Y.*

Quasi-experimental designs (QEDs): Used to determine whether an intervention caused the intended outcomes. In QEDs individuals are not randomly assigned to groups because of ethical or practical constraints. Instead, equivalent groups are created through matching or other statistical adjustments.

Randomized controlled trials (RCTs): Used to determine whether an intervention caused the intended outcomes. RCTs involve randomization, a process like a coin toss, to assign individuals to the treatment or comparison group.

Treatment group: The group that receives the intervention.

Comparison group: The group that does not receive the intervention and is used as the counterfactual to the intervention.

Validity: The extent to which the results of an evaluation are supportable, given the evaluation design and the methods used. Validity applies to the evaluation design, analytic methods, and data collection. Ultimately, valid claims are sound ones. There are two main types of validity:

Internal validity: The extent to which a study or instrument measures a construct accurately and is free of alternative explanations. There are two common threats to internal validity:

Attrition: When participants (individuals, schools, and so on) leave an evaluation before it concludes.

Selection bias: When the treatment group differs from the comparison group in a meaningful way that is related to the outcomes of interest.

External validity: The extent to which an instrument or evaluation findings can be generalized to different contexts, such as other populations or settings. There are three common threats to external validity:

Contextual factors of populations: When contextual factors, such as time and place, differ between the sample in the evaluation and a population to which one wants to generalize.

Multiple treatments: When external factors, such as an additional program, might cause the evaluation to detect a different effect than it would if the external factors were not present.

Hawthorne effect: When individuals act differently because they are aware that they are taking part in an evaluation.

Evidence-based programs: Programs that have evidence of their effectiveness in producing results and improving outcomes when implemented.

The Every Student Succeeds Act (ESSA): A law that encourages state and local education agencies to use evidence-based programs. There are four ESSA tiers of evidence (U.S. Department of Education, 2016). These tiers fall under the Education Department General Administrative Regulations Levels of Evidence for research and evaluation design standards:

Strong evidence: A program is supported by at least one well-implemented randomized controlled trial with low attrition. Attrition refers to the number of participants who leave a study before it is completed.

Moderate evidence: A program is supported by at least one well-implemented randomized controlled trial with high attrition or at least one well-implemented quasi-experimental design.

Promising evidence: A program is supported by at least one well-implemented correlational design with statistical control for selection bias.

Demonstrates a rationale: A program has a well-specified logic model with one intended outcome of interest that aligns with a stakeholder need. The program is supported by existing or ongoing research demonstrating how it is likely to improve the outcomes identified in the logic model.

What Works Clearinghouse (WWC) design standards: The WWC is part of the U.S. Department of Education's Institute of Education Sciences. To provide educators with the information they need to make evidence-based decisions, the WWC reviews research on education programs, summarizes the findings of that research, and assigns evidence ratings to individual studies (What Works Clearinghouse, 2020).

There are three WWC design standards that correspond to the ESSA tiers of evidence. A study can be found:

To meet WWC standards without reservations: This tier corresponds to strong evidence under ESSA.

To meet WWC standards with reservations: This tier corresponds to moderate evidence under ESSA

Not to meet WWC standards: This tier still provides promising evidence under ESSA.

Module 4: Evaluation samples

Population: All possible participants in a program.

Census: Used to collect data from everyone in a population.

Sample: A subset of an entire population that is identified for data collection.

Representativeness: How well a sample represents the entire population.

Generalizability: The extent to which the results of an evaluation apply to different types of individuals and contexts.

Weighting: Statistical adjustments to ensure a sample is representative of the entire population with respect to particular characteristics.

Sample size: The number of participants needed in a sample to collect enough data to answer the evaluation questions.

Sampling frame: A list of all possible units (such as students enrolled in schools in a particular district) that can be sampled.

Random sampling: A sampling technique in which every individual within a population has a chance of being selected for the sample. There are three main types of random sampling:

> **Simple random sampling:** Individuals in a population are selected with equal probabilities and without regard to any other characteristics.
>
> **Stratified random sampling:** Individuals are first divided into groups based on known characteristics (such as gender or race/ethnicity). Then, separate random samples are taken from each group.
>
> **Clustered random sampling:** Individuals are placed into specific groups, and these groups are randomly selected to be in the sample. Individuals cannot be in the sample if their groups are not selected.

Nonrandom sampling: A sampling technique in which only some individuals have a chance of being selected for the sample. There are four main types of nonrandom sampling:

> **Consecutive sampling:** Individuals meeting a criterion for eligibility (such as being math teachers) are recruited until the desired sample size is reached.
>
> **Convenience sampling:** Individuals are selected who are readily available and from whom data can be easily collected.

Snowball sampling: Individuals are recruited through referrals from other participants.

Purposive sampling: Individuals are selected to ensure that certain characteristics are represented in the sample to meet the objectives of the evaluation.

Saturation: The point at which the data collected begin to yield no new information and data collection can be stopped.

Unit of measurement: The level at which data are collected (for example, student, classroom, school).

Confidence interval: A range of values for which there is a certain level of confidence that the true value for the population lies within it. The range of values will be wider or narrower depending on the desired level of confidence. Standard practice is to use a 95 percent confidence level, which means there is a 95 percent chance that the range of values contains the true value for the population.

Null hypothesis: A statement that suggests there will be no difference between the treatment group and the comparison group involved in an evaluation.

Statistical power: The probability of rejecting the null hypothesis when a particular alternative hypothesis is true.

Continuous data: Data that can take on a full range of possible values, such as student test scores, years of teaching experience, and schoolwide percentage of students eligible for the National School Lunch Program.

Binary data: Data that can take on only two values (yes or no), such as pass or fail scores on an exam, course completion, graduation, or college acceptance.

Standard deviation: A measure that indicates how spread out data are within a given sample.

Module 5: Data quality

Quantitative data: Numerically measurable information, including survey responses, assessment results, and sample characteristics such as age, years of experience, and qualifications.

Qualitative data: Information that cannot be measured numerically, including interview responses, focus group responses, and notes from observations.

Data quality: The extent to which data accurately and precisely capture the concepts they are intended to measure.

Validity: The extent to which an evaluation or instrument really measures what it is intended to measure. Validity applies to the evaluation design, methods, and data collection. There are two main types of validity:

> **Internal validity:** The extent to which a study or instrument measures a construct accurately and is free of alternative explanations.
>
> **External validity:** The extent to which an instrument or evaluation findings can be generalized to different contexts, such as other populations or settings.

Reliability: The extent to which the data source yields consistent results. There are three common types of reliability:

> **Internal consistency:** The extent to which items in a scale or instrument consistently measure the same topic.
>
> **Test–retest reliability:** The extent to which the same individual would receive the same score on repeated administrations of an instrument.
>
> **Inter-rater reliability:** The extent to which multiple raters or observers are consistent in coding or scoring.

Timeliness: The extent to which data are current and the results of data analysis and interpretation are available when needed.

Comprehensiveness: The data collected in an evaluation include sufficient details or contextual information and can therefore be meaningfully interpreted.

Trustworthiness: The extent to which data are free from manipulation and entry error. Trustworthiness is often addressed by training data collectors.

Completeness: Data are collected from all participants in the sample and are sufficient to answer the evaluation questions. Completeness also relates to the degree of missing data and the generalizability of the dataset to other contexts.

Triangulation: Reviewing multiple sources of data to look for similarities and differences.

Member checks: Establishing the validity of qualitative findings through key stakeholder and participant review.

Audit trail: A documented history of qualitative data collection and analysis. Careful documentation of data collection procedures, training of data collectors, and notes allows for findings to be cross-referenced with the conditions under which the data were collected.

Evaluation matrix: A planning tool to ensure that all necessary data are collected to answer the evaluation questions.

Module 6: Data collection

Interview: Directly asking an individual questions to collect data to answer an evaluation question.

Focus group: Directly asking a group of participants questions to collect data to answer an evaluation question.

Protocol: Instructions for conducting an interview, focus group, or observation. An interview or focus group protocol should include steps for conducting the interview or focus group, a script of what to say, and a complete set of questions. An observation protocol should include information about items to observe, the data collection approach to use (recording checklist, observation guide, or open field notes), and the type of observation.

Observation: Watching individuals or groups to collect information about processes, situations, interactions, behaviors, physical environments, or characteristics. There are four types of observation, all of which can be conducted in person or virtually:

> **Controlled observation:** Conducted in structured and arranged settings.
>
> **Natural observation:** Conducted in unstructured and real-life settings.
>
> **Overt observation:** Observers make their presence known.
>
> **Covert observation:** Observers do not make their presence known.

Survey: Administering a fixed set of questions to collect data in a short period. Surveys can be an inexpensive way to collect data on the characteristics of a sample in an evaluation, including behaviors, practices, skills, goals, intentions, aspirations, and perceptions.

Observable variable: Behaviors, practices, or skills that can be directly seen and measured. Also called a measurable variable. These data are collected in a variety of ways (for example, observations, surveys, interviews).

Unobservable variable: Goals; intentions; aspirations; or perceptions of knowledge, skills, or behavior that cannot be directly seen and measured but can be inferred from observable indicators or self-report. Also called a latent variable.

Open-ended question: A question that does not include fixed responses or scales but allows respondents to add information in their own words.

Close-ended question: A question that includes fixed responses such as yes or no, true or false, multiple choice, multiple selection, or rating scales.

Midpoint: The middle of a rating scale with an odd number of response options. Typically, respondents can select the midpoint to remain neutral or undecided on a question.

Double-barreled question: A question that asks two questions but forces respondents to provide only one answer. For example, "Was the professional development culturally and developmentally appropriate?"

Loaded question: A question that could lead respondents to answer in a way that does not represent their actual position on the topic or issue. For example, the wording of a question or its response options could suggest to respondents that a certain answer is correct or desirable.

Probing question: A follow-up question that helps gain more context about a particular response or helps participants think further about how to respond.

Recording checklist: A standardized form, with preset questions and responses, for observing specific behaviors or processes.

Observation guide: A form that lists behaviors or processes to observe, with space to record open-ended data.

Open field notes: A flexible way to document observations in narrative form.

Mutually exclusive: When two response options in a survey cannot be true at the same time.

Collectively exhaustive: When response options in a survey include all possible responses to a question.

Module 7: Data analysis

Data preparation: Collecting, organizing, and cleaning data in a manner that ensures accurate and reliable analysis.

Data error: The difference between an actual data value and the reported data value.

Outlier: A data value that is positioned an abnormal distance from the expected data range.

Data analysis: The process of examining and interpreting data to answer questions. There are two broad approaches to data analysis:

> **Descriptive methods:** Describing or summarizing a sample. Descriptive methods can involve examining counts or percentages; looking at the central tendency of a distribution through means, medians, or modes; and using statistics such as standard deviation or interquartile range to look at the spread, or variation, of a distribution.
>
> **Inferential methods:** Drawing conclusions about a population from a sample. Inferential methods can include techniques such as *t*-tests, analysis of variance (ANOVA), correlation, and regression.

Mean: The average response across a sample.

Median: The value at the midpoint of a distribution.

Mode: The most common response in a distribution.

Standard deviation: A measure of how spread out data points are that describes how far the data are from the mean.

Range: The maximum and minimum observed values for a given variable.

Quartile: One of four even segments that divide up the range of values in a dataset.

Interquartile range: The spread of values between the 25th percentile and the 75th percentile.

***t*-test:** A comparison of two means or standard deviations to determine whether they differ from each other.

Analysis of variance (ANOVA): A comparison of three or more means that determines whether there are statistically significant differences among them.

Correlation analysis: Analysis that generates correlation coefficients that indicate how differences in one variable correspond to differences in another. A positive correlation

coefficient indicates that the two variables either increase or decrease together. A negative correlation coefficient indicates that, as one variable increases, the other decreases.

Regression analysis: A family of statistical procedures that estimate relationships between variables.

> **Simple or linear regression analysis:** Analysis that can show the relationship between two variables.
>
> **Multiple regression analysis:** Analysis that can control for other factors by including additional variables.

Dependent variable: A variable that could be predicted or caused by one or more other variables.

Independent variable: A variable that has an influence on or association with the dependent variable.

Covariate: A variable that has a relationship to the dependent variable that should be considered but that is not directly related to a program. Examples of covariates are student race/ethnicity, gender, socioeconomic status, and prior achievement.

Confound: A variable that could result in misleading interpretations of a relationship between the independent and dependent variable. For example, if all the teachers who are implementing a new math intervention program have a master's degree in math while the teachers who are not implementing the program have only a bachelor's degree, the degree attainment of the intervention teachers is a confound. Teachers' additional education experience, rather than the math intervention, could be the reason for changes in student achievement.

Module 8: Dissemination approaches

Dissemination: Sharing information about an evaluation and its findings with a wide audience.

Dissemination plan: Strategically planning dissemination activities to use time and other resources efficiently and to communicate effectively.

Audience: The group of people who need or want to hear the information that will be disseminated.

Message: The information that the audience needs to know about an evaluation and that the evaluators want to share.

Approach: The means used to disseminate the information to the audience. There are many dissemination approaches:

Blog: An online forum for sharing regular updates about a program and the evaluation process.

Data dashboard: A visual tool for organizing and sharing summaries of large amounts of data, especially quantitative data.

In-person meeting: A gathering of interested stakeholders at which an evaluator presents the findings through multimedia and visual displays of the data.

Media release: A write-up about an evaluation and its findings to be shared with media outlets.

Evaluation report: A formal, highly organized document describing the methods, measures, and findings of an evaluation.

Evaluation brief: A condensed version of an evaluation report that provides a brief overview of the methods and findings.

Summary of findings: A short one- to two-paragraph piece that briefly describes what is happening and what was found in an evaluation.

Social media: Digital tools to quickly create and share information about an evaluation with a variety of audiences.

Webinar: A visual medium and way to reach large numbers of people, often at little or no cost.

Infographic: A one- or two-page document that graphically represents data and findings to tell a story.

Video: A way to share information quickly, clearly, and in an engaging way.

Podcast: A brief recording for sharing information on a topic through a discussion format.

Timing: When the audience needs to know the information.

Plain language: Using clear communication and writing so that it is easy for the audience to understand and use the findings of an evaluation.

Accessibility: Ensuring that dissemination products are available to all individuals, including people with disabilities, by meeting the requirements for Section 508 compliance.

Data visualization: Using graphical representations so that data are easier to understand.

Alternative text: A narrative description of a figure, illustration, or graphic for readers who might not be able to engage with the content in a visual form.

References

U.S. Department of Education. (2016). *Non-regulatory guidance: Using evidence to strengthen education investments.* https://www2.ed.gov/policy/elsec/leg/essa/guidanceuseseinvestment.pdf.

What Works Clearinghouse. (2020). *Standards handbook* (Version 4.1). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance. https://ies.ed.gov/ncee/wwc/handbooks.

Acknowledgments

The Program Evaluation Toolkit would not have been possible without the support and contributions of Trudy Cherasaro, Mike Siebersma, Charles Harding, Abby Laib, Joseph Boven, David Alexandro, and Nazanin Mohajeri-Nelson and her team at the Colorado Department of Education.

REL 2022–112

October 2021

This resource was prepared for the Institute of Education Sciences (IES) under Contract ED-IES-17-C-0005 by the Regional Educational Laboratory Central administered by Marzano Research. The content of the resource does not necessarily reflect the views or policies of IES or the U.S. Department of Education, nor does mention of trade names, commercial products, or organizations imply endorsement by the U.S. Government.

Stewart, J., Joyce, J., Haines, M., Yanoski, D., Gagnon, D., Luke, K., Rhoads, C., & Germeroth, C. (2021). *Program Evaluation Toolkit: Quick Start Guide* (REL 2022–112). U.S. Department of Education, Institute of Education Sciences, National Center for Education Evaluation and Regional Assistance, Regional Educational Laboratory Central. http://ies.ed.gov/ncee/edlabs.

This resource is available on the Regional Educational Laboratory website at http://ies.ed.gov/ncee/edlabs.