RAND School Leadership Intervention Evaluation Toolkit

Stephani L. Wrabel, Rebecca Herman, Susan M. Gates

January 31, 2018
Prepared for the Wallace Foundation
**Preface**

The reauthorization of the U.S. Elementary and Secondary Education Act, referred to as the Every Student Succeeds Act (ESSA), emphasizes evidence-based interventions while providing new flexibility to states and districts with regard to the use of federal funds, including funds to promote effective school leadership. This new flexibility comes with an expectation that educators will adopt evidence-based improvement activities. However, in a change under ESSA, educators now have the option to implement activities that have weaker evidence, if there is ongoing evaluation of the impact. For example, a district might adopt a newly developed professional development program for principals, using resources authorized in ESSA, even when the program does not have established evidence of effects. This is permissible so long as the program is based on a logic model and an evaluation of the program occurs.

Critically examining program implementation and outcomes can be useful for educators. Evaluations can help state and local policymakers and administrators make decisions about whether to sustain or scale up existing efforts. This toolkit aims to help those responsible for school leadership interventions examine the effects of their interventions and responsibly allocate scarce resources. In particular, we expect it to be of interest to state and local (e.g., county, district) education administrators seeking to leverage federal funding to support leadership programs that meet ESSA requirements as they design evaluations of those programs.

The toolkit builds on information presented in a companion document, *School Leadership Interventions Under the Every Student Succeeds Act: An Evidence Review*, which describes the opportunities ESSA provides for supporting school leadership, discusses the standards of evidence delineated under ESSA, and synthesizes the research base with respect to those standards.

The toolkit was modeled after the design and program evaluation tools of the RAND Suicide Prevention Program Evaluation Toolkit developed by Joie D. Acosta, Rajeev Ramchand, Amy Becker, Alexandria Felton, and Aaron Kofner. This toolkit was selected as a premier resource for the Suicide Prevention Resource Center.

This research was conducted by RAND Education, a division of the RAND Corporation, with grant funding from The Wallace Foundation. The Wallace Foundation is committed to improving school leadership through better training, hiring, support, and evaluation of principals. For more than a decade, it has invested in research, initiatives, and evaluations to improve school and district leadership and contribute to an evidence base in this area.
Text content is not provided.
Executive Summary
This toolkit is a companion to School Leadership Interventions Under the Every Student Succeeds Act: An Evidence Review (the School Leadership Interventions report), which assessed the evidence on school leadership interventions against Every Student Succeeds Act (ESSA) evidence tiers. An outcome evaluation is an ESSA requirement for federally funded interventions that have low levels of evidence (e.g., Tier IV). To address this, we developed this toolkit to help state and local education policymakers evaluate school leadership intervention outcomes as they implement the interventions. This toolkit is designed primarily for administrators working in state education agencies (SEA) and local (e.g., county, district) education agencies (LEA) considering school leadership improvement interventions as a school improvement lever. We provide guidance and tools for building an intervention logic model, selecting program evaluation designs and measures, organizing and analyzing data, and using data to improve the intervention. The guidance shared in this toolkit is consistent with the larger body of evaluation methodology but is specifically tailored to be relevant for school leadership intervention evaluations. Although the toolkit focuses on outcome evaluations, we strongly encourage state and local education administrators to go beyond the ESSA requirements to look at implementation processes as well (see Appendix B).
Chapter 1: Introduction and Overview

Principals have a powerful influence on the educational outcomes of students in U.S. schools (Leithwood et al., 2004; Coelli and Green, 2012; Dhuey and Smith, 2014; Grissom, Kalogrides, and Loeb, 2015). There is growing recognition among educators and education policymakers around the country that efforts to improve school leadership practices and the school leadership pipeline have the potential to improve student outcomes. There is also an underlying understanding that practices to improve outcomes should be based on evidence. In fact, the 2015 reauthorization of the Elementary and Secondary Education Act, referred to as the Every Student Succeeds Act (ESSA), sets the expectation that education activities supported under ESSA—including efforts to strengthen school leadership—should have demonstrated evidence of success.

This toolkit is a companion to School Leadership Interventions Under the Every Student Succeeds Act: An Evidence Review (the School Leadership Interventions report). School Leadership Interventions looked at six types of school leadership interventions: leader evaluation systems, principal preparation programs, strategic staff management, professional learning, working conditions, and leadership-focused school reform models. Periodically throughout this report, we provide examples using one or more of these types of interventions. These six intervention types are described in more detail in the “Tip for Getting Started: What Type of Program Do I Have” information box at the end of this chapter, and specific interventions reviewed in the original report are described in Appendix D of School Leadership Interventions.

The School Leadership Interventions report assessed the evidence on school leadership interventions within these six categories against ESSA evidence tiers. According to ESSA, an evidence-based activity, strategy, or intervention must show statistically significant positive effects on student or other relevant outcomes (e.g., staff turnover, school climate), based on one or more of the following (Pub. L. 114-95, 2015, §8101):

- Tier I (strong evidence)—at least one well-designed and well-implemented experimental study
- Tier II (moderate evidence)—at least one well-designed and well-implemented quasi-experimental study
- Tier III (promising evidence)—at least one well-designed and well-implemented correlational study that controls for selection bias.

For federally funded activities other than Title I school-improvement activities—which includes many of the funding streams to support school leadership activities—Tier IV also is considered sufficient evidence:

- Tier IV—the activity, strategy, or intervention demonstrates a rationale based on high-quality research or a positive evaluation that suggests it is likely to improve student or other relevant outcomes. For Tier IV activities, there must be ongoing efforts to evaluate the effects of the activity, strategy, or intervention.

Although the review found many leadership interventions with Tiers I through III evidence, there are many others that would qualify only for Tier IV. Under ESSA, SEA and LEA administrators are charged with identifying effective programs on the basis of existing evidence and, when they select programs without Tiers I through III evidence, ensuring that their school–leadership improvement program is being

---

1 We encourage users of this toolkit to review the School Leadership Interventions report to gain a clearer understanding of the evidence (Herman et al., 2016).

2 Note that ESSA does not specify the nature of the evaluation efforts, and it is not clear that such an evaluation will lead to evidence that qualifies as Tiers I, II, or III.
evaluated. We developed this toolkit to help state and local (e.g., district) education policymakers evaluate school leadership interventions as they implement the interventions.³

Not all evidence is created equal—while ESSA allows more- and less-rigorous studies to be used to justify adopting an intervention, more-rigorous evidence provides a better indication of the impact of the intervention. The U.S. Department of Education’s nonregulatory guidance on ESSA’s evidence requirements recommends prioritizing strong and moderate evidence (U.S. Department of Education, 2016). For those designing evaluations to support the use of school-leadership interventions, we encourage the use of the most rigorous research design feasible to provide the best information for education decisionmakers. In Chapter 3 of this toolkit, we link specific program evaluation designs to the ESSA tier of evidence they have the potential to produce. Although the toolkit focuses on impact evaluations, we strongly encourage SEA and LEA administrators to go beyond the ESSA requirements to also look at implementation processes as introduced in Chapter 3 and explored in Appendix B.

Whom Is This Toolkit for?
This toolkit is designed primarily for administrators working in SEAs and LEAs considering school leadership improvement interventions as a school improvement lever. We hope this toolkit will help such administrators evaluate school leadership improvement programs as they are implementing the programs.⁴ The toolkit lays out larger concepts to engage administrators in thinking about the evaluation, recognizing that in-house or external researchers may support administrators with the evaluation. While the information within this toolkit focuses on school leadership interventions, the information presented may be adapted for additional areas of educational program evaluation.

Toolkit Goals
The overarching goal of this toolkit is to help administrators responsible for implementing school leadership interventions evaluate those efforts to determine whether they are achieving desired outcomes. This toolkit provides

1. insight about the latest evaluation research in general and that associated with school leadership interventions specifically
2. guidance to select an appropriate evaluation approach based on the type of intervention chosen and the resources and expertise available for the evaluation
3. support for new evaluation efforts or additional information to advance evaluation efforts.

Each chapter of the toolkit contributes to these specific aims, which can be achieved by working through the entire toolkit. The toolkit follows the flow of an evaluation. We begin with conceptualizing the intervention and then explore evaluation designs that fit the intervention type. We then turn to data and measures, analysis, and reporting. Throughout, we provide examples to illustrate how the evaluation might look for different types of principal interventions.

✔ Chapter 2 provides guidance to identify the core components of a school leadership program and organize these components into a logic model to clearly illustrate the relationships and

³ School leadership improvement activities, strategies, and interventions are designed to improve school leadership, changing leaders’ behaviors with the intent of ultimately improving student outcomes. We use the terms “activities,” “strategies,” “interventions,” and “initiatives” somewhat interchangeably, although there are differences among them. School leadership improvement interventions are designed to act upon the principal. We do not consider school leadership behaviors as interventions; rather, behaviors are the result of interventions.

⁴ Educators selecting among existing programs should consider the existing evidence (see the School Leadership Interventions report) and might also consider aspects of the context in which the program would operate (see the Hexagon Tool [Blasé, Kiser, and Van Dyke, 2013] and other resources of the National Implementation Research Network [Blasé, Van Dyke, and Fixsen, 2013]).
dependencies between components. The chapter also contains a tool for reviewing a logic model and assessing whether it is complete and reasonable.

- Chapter 3 provides information on different types of evaluations and guidance on how to select the right method to evaluate your program.
- Chapter 4 provides guidance on selecting the right measures for your evaluation, including a review of different types of data and key indicators.
- Chapter 5 helps you organize and analyze data about your program.
- Chapter 6 helps you assess your evaluation and interpret evaluation results to identify next steps based on what you learned from the analysis of your program data.

Fully completing each chapter will help you ensure that you have taken advantage of the comprehensive guidance provided. However, we recognize that SEA and LEA administrators may come to this process with different levels of prior knowledge and encourage users to incorporate their own evaluation expertise as they reflect on the guidance in the toolkit.

**Evaluation Challenges Addressed**

Program evaluation does not come without its own set of challenges, some of which are universal to evaluation efforts and some more specific to the evaluation of school leadership interventions. Improvement in one or more student outcomes is the ultimate objective of all education improvement interventions. Burkhauser and colleagues (2012) provide an in-depth discussion of the challenges associated with using student outcome data to evaluate the effectiveness or impact of interventions targeting school leaders. But evaluation challenges extend beyond data issues. Next, we summarize key evaluation challenges and describe practical solutions for addressing them, linking those solutions to resources in this toolkit.

**Challenge 1: Many school leadership interventions have multiple components. It can be difficult to attribute improved outcomes to a specific component or characteristic of the chosen intervention.**

Developing a logic model, as described in Chapter 2 of this toolkit, helps tie specific activities of a multifaceted intervention with intended outcomes. Understanding the connections between activities and outcomes provides decisionmakers with the information needed for continuous improvement of an intervention or selection of future interventions. Assessing the implementation of each component of the intervention can help identify where interventions meet or fail to meet expectations to provide additional insight and information needed for continuous improvement and future policy responses. (See Appendix B.)

**Challenge 2: The effect of school leadership interventions on student outcomes may not be measurable for many years.** Program evaluations can examine short-term, intermediate, and long-term outcomes. Short-term outcomes may provide implementers with useful feedback to guide early decisionmaking about program delivery or operations. (see Chapter 4). Short-term and intermediate outcomes can include measures other than student outcomes. This interim information can help decisionmakers see whether an intervention is on track to improve student outcomes.

**Challenge 3: Evaluators must account for school leader placement:** To influence student outcomes, a school leadership program must “treat” an individual who is currently serving or will serve as a school leader. Evaluators of programs targeting future school leaders (such as preservice programs) must account for the principal placement process. Not all program participants will become school leaders, and preservice program participants may not be placed as a principal for many years after completing the program, thus limiting the feasibility of using an experimental design (see Chapter 3).

**Challenge 4: Schools serve diverse communities of students, and differences in student background may artificially inflate or deflate estimates of program impact.** Chapter 3 of
this toolkit outlines common program evaluation methods. Some of these methods are able to address the nonrandom sorting of students across schools to closely identify the actual effect of a program on outcomes rather than the influence of student characteristics on those outcomes.

**Challenge 5: School context can mediate the relationship between a chosen school leadership intervention and its impact.** Context may also affect the time it takes for desired impacts to be realized. This challenge must be considered in selecting an appropriate evidence-based intervention and when designing an evaluation. When selecting an intervention, consider whether existing evidence is based on studies involving a similar population or setting (see Chapter 2). When designing an evaluation, consider options for accounting for context. For example, control group and comparison group analyses that control for initial differences can account for some of these contextual differences.

**Challenge 6: Some types of school leadership interventions are difficult to evaluate in isolation.** For example, school districts generally make the decision to replace a principal from a low-performing school (strategic staff management) based on a number of factors, such as school performance, contractual arrangements, parental and community pressure, and past performance of the principal. It would be very difficult to design a rigorous study that focuses only on the impact of principal replacement without all the other factors affecting the findings. Chapter 3 suggests the appropriateness of different evaluation designs for different types of school leadership interventions.

**How the Toolkit Was Developed**
To create this toolkit, we conducted a review of the program evaluation literature, including existing toolkits and guidance on program evaluation in education and related areas. The toolkits reviewed include the *Getting to Outcomes™ Guide for Teen Pregnancy Prevention* (Chinman et al., 2016), the *Suicide Prevention Program Evaluation Toolkit* (Acosta et al., 2013), *Getting to Outcomes™ 2004 Promoting Accountability Through Methods and Tools for Planning, Implementation, and Evaluation* (Chinman, Imm, and Wandersman, 2004), and the W.K. Kellogg Foundation’s (2004a) *Evaluation Handbook*. This toolkit draws extensively from the *Suicide Prevention Program Evaluation Toolkit*, which was selected as a premier resource for the Suicide Prevention Resource Center. From these materials, we identified key steps in the evaluation process and strategies for presenting the information clearly.

The school leadership–related information presented in this toolkit comes from the *School Leadership Interventions* report. That report was based on a systematic review of literature. Additional details on the review method and data sources can be found in Herman et al. (2017).
Tip for Getting Started: What Kind of Program Do I Have

Before beginning, we encourage toolkit users to identify which type of program they will be implementing. Understanding how similar types of interventions have been evaluated may help users identify an appropriate evaluation method, develop a logic model, or select the types of measures and data for evaluating a similar intervention. In the School Leadership Interventions report, we identify six types of leadership programs as having established evidence that meets at least one of the four ESSA evidence tiers. School reform models must have evidence in Tiers I through III to receive federal funds. The other five types of programs can use evidence from Tiers I through IV. If your program falls into more than one of the broad categories, consider how each type of intervention is evaluated.

In this box, we provide a broad synopsis of each of these program types. More information on our review of the evidence related to these program categories is available in Appendix A, as well as in the School Leadership Interventions report.

Leader evaluation systems are a set of processes, tools, and metrics designed to evaluate principals’ strengths and needs—for either accountability or developmental purposes. They may be aligned with rigorous leadership standards and are often developed by the state or district education agency, sometimes drawing on nationally available tools.

Principal preparation programs, broadly defined, involve a classroom-based graduate school program and some type of school-based internship and can lead to an advanced degree or certification for an aspiring principal. They may be provided by universities, districts, or independent organizations, or some combination of the three. ESSA defines principal preparation programs as operated by a public or other nonprofit organization (including or affiliated with an institution of higher education), containing a clinical preparation course (where the student is paired with an effective educator) and instruction in content areas, committed to producing a specified number of effective educators, and requiring demonstrated effectiveness to award a certificate or degree.

Strategic staff management may include activities to improve recruitment and selection processes, placement of principals in schools, and principal replacement. Recruitment and retention interventions may include, for example, communication strategies to broaden the candidate pool or specialized processes and tools to screen and evaluate candidates (e.g., performance-based interview tasks).

Professional learning includes a variety of learning experiences for sitting school principals, such as professional development through workshops (single sessions or a series) and coaching or mentoring. These opportunities may be available throughout the principal’s career, although they often are most intensive early in his or her career or placement at a school. The focus of these learning programs is often on developing competencies related to job responsibilities (e.g., managing the change process, setting vision, hiring and supervising staff).

Working conditions can include supports and policies designed specifically to improve school leader effectiveness, such as school autonomy and performance incentives. There are many other working conditions (e.g., school climate) that likely mediate or moderate leaders’ effectiveness but are not the focus of the intervention.

Leadership-focused school reform models are multidimensional activities (e.g., changes in curriculum, instruction, staffing, management) focused on improving low-performing schools. For this toolkit, we focus on comprehensive school reform models that prioritize school leadership improvement as one of a small number of core components.
Chapter 2: Building a Program Logic Model

It may seem to educators that school leadership interventions are distant from students and classrooms and unlikely to have an effect. School leaders are not in the classroom working with children day in and day out. However, research shows that school leaders do affect student outcomes. To build support for school leadership interventions, SEA and LEA administrators must be able to link the intervention to what happens in the classroom and monitor whether interventions are working as planned.

The first step to accomplishing these goals is developing a logic model that links the intervention’s activities to the outcomes that stakeholders care about through a clear chain of events. Usually, logic models are a visual representation of the hypotheses and assumptions regarding how an intervention, if implemented as planned, creates change within and helps achieve desired outcomes of an organization (i.e., a school) (McLaughlin and Jordan, 2010). They typically model relationships using boxes and arrows, but the key connections can be communicated in a variety of modes. A logic model can help administrators justify and build support for an intervention and guide the evaluation plan. It can help evaluators identify the outcomes that should be the focus of the intervention. For administrators looking to select an intervention that already has established evidence, a logic model can help identify the resources necessary for the intervention to be successful, the types of changes a particular intervention is expected to produce, and where rigorous evidence already exists in support of the program. Finally, a logic model based on prior evidence of effectiveness can increase confidence that a new intervention is likely to produce the desired outcomes. In fact, the ESSA evidence requirements, at Tier IV, require a logic model based on prior research, as well as confirmation that evaluation is being conducted on the intervention (through or outside of the project in question) for funding some types of school leadership improvement efforts.

In this chapter, we discuss the critical components of a logic model and provide a logic model template to guide your work. We also provide suggestions for additional resources and materials that offer further guidance and support.

Developing a Logic Model

A team of SEA and LEA administrators, program stakeholders, and other knowledgeable staff should work collaboratively to ensure that the school leadership program logic model is accurate and complete. The process of building the logic model together helps the team members check their assumptions and reach a common understanding of the intervention. A well-specified logic model helps with the development of the evaluation plan (see Chapter 3).

As noted at the end of the previous chapter, you may wish to consult studies evaluating programs similar to the one you have chosen to get examples of logic models that may help you develop your own.

Core Components

Below is a brief introduction to each component in a logic model. We have provided a few questions for administrators to consider as they gather the necessary information. These questions are intended to help identify key aspects of how a program operates. Importantly, these are suggested questions rather than an exhaustive list and will differ by the type of intervention selected.
RESOURCES are the inputs needed and available to support the program.
1. What amount of funding will pay for the intervention, and where does that money come from (e.g., staff training and professional development funds or grants, tuition or registration funds)? How does the time line for funding affect intervention planning and implementation?
2. What staff will be involved in delivering the program? What staff will provide support to the implementers, and what will be the nature of that support? Is turnover in staff likely to affect implementation?
3. What facilities will be available (e.g., staff conference room, classrooms), or what technology will be utilized in the program (e.g., computers for webinars)?
4. What amount of time must principals, school leaders, or other staff commit for the program? When in the year should they be most actively engaged?

For example, consider leader evaluation systems. Resources should be budgeted to develop the evaluation system or tailor it (if an existing system is used), support training, and cover time spent using the new system. District staff will likely be needed to develop or tailor the system, develop and deliver training for school leaders, evaluate leaders, and provide feedback to leaders. Facilities might include an online or paper-based platform for the evaluation system, as well as training space. Some of these resources may come from existing district budget items. The district might also apply for state or federal funding to support the new system.

ACTIVITIES are processes, tools, events, and actions that constitute the school leadership intervention.
1. What events, trainings, activities, and meetings will be held, and when? How many of each will participants attend?
2. What communication will be sent to staff (e.g., staff will be sent weekly reminders to fill out activity surveys)?
3. Will supports be individualized for school leaders or delivered consistently across participants?

For example, leader evaluation system activities will likely include development, leader training, evaluation, and feedback. Depending on the specific leader evaluation system, there might be several training events—perhaps one on the evaluation process and one on the feedback process. There might be multiple evaluation events—perhaps quarterly evaluation visits to each school. Communication to principals might include a description of the new system and processes, reminders to complete activities, and evaluation feedback. Perhaps some principals (e.g., early career principals) might participate more intensively than others. Again, the specific activities depend heavily on the specific leader evaluation system and how it is used in the district.

OUTPUTS are the observable products, goods, and services delivered during the program, regardless of whether the desired or intended outcomes are met.
1. How many staff will participate in some aspect of the intervention? How many or what proportion of schools do these staff members serve?
2. How many hours of training or development will be provided?

For example, outputs of a leader evaluation system might be that 100 percent of school leaders participated in all training and evaluations, and a professional development plan was developed for every school leader.

OUTCOMES are the aspects of the leader, the leader’s work environment (e.g., staff, school), or the leader’s constituents that are expected to change as a result of the program. Outcomes should be SMART (Doran, 1981):

- Specific
- Measurable
- Action-oriented (focused on tasks or actions controlled by the SEA or LEA)
Traditionally, logic models separate outcomes into three categories: short term, intermediate, and long term. While these categories may relate to the time needed to see the results (for example, short-term outcomes might occur within a year, and long-term outcomes might occur several years later), they also suggest that outcomes of the intervention may be sequential or additive (McLaughlin and Jordan, 2010). Long-term outcomes result from the short-term and/or intermediate outcomes brought about by the program. In the case of principals, we might think of short-term outcomes being the changes in the principal’s actions; intermediate outcomes as the changes that the principal directly affects in the overall school environment (e.g., school climate), staffing, or instruction; and long-term outcomes as being student-level improvements.

Student achievement is an example of a long-term outcome that might be expected to change following many of the school leader interventions (e.g., professional learning, leadership-focused school reform models). For student achievement to be a SMART outcome, evaluators must explicitly define the measure used and the benchmark (where applicable) against which the measure will be compared. For example, student achievement can be measured by standardized reading test scores of students in all tested grades within a school. Evaluators can compare the average achievement of pre- and postintervention cohorts of students served by the school leader (i.e., measuring improvement over time) or against a preestablished benchmark (e.g., 0.15–standard deviation improvement) (more on establishing appropriate benchmarks in Chapter 4). Evaluators should also establish how long after an intervention the desired level of change is expected. In the end, this SMART outcome could be phrased as “An average of 0.15 standard deviations improvement in standardized reading test scores between preintervention and postintervention cohorts of students in the leader’s school within four years of intervention start.” If the measure will not be compared with a benchmark but will be evaluated against a nonintervention control or comparison group, the SMART outcome might be written as “The average improvement in standardized reading test scores between preintervention and postintervention cohorts of students within four years of intervention start.”

Extending this example, leader evaluation systems might have short-term outcomes (e.g., principals conduct 10 percent more classroom observations and teacher follow-up focused on questioning strategies), intermediate outcomes (e.g., teachers ask a mix of lower- and higher-level questions in each discussion period), and long-term outcomes (e.g., students’ comprehension significantly improves as measured by standardized tests).

When developing outcomes for the logic model, think through the following questions:
1. What principal characteristics, behaviors, or outcomes does the intervention aim to change?
2. How will changes in the principal lead to changes in school climate, instruction, staffing, and student outcomes? How should instruction change?
3. What student outcomes are expected to change as a result of the changes to principal and teacher skills (e.g., absenteeism, test scores, engagement)?
4. When should these outcomes be measured?

Logic models can also contain a section on local and external CONTEXT. The program may work differently in different settings. The logic model can identify the most critical contextual factors that

---

5 For this toolkit, we use the first 12 months, one to three years, and four to six years for the time frame of short-term, intermediate, and long-term outcomes. The W. K. Kellogg Foundation uses one to three years as short-term, four to six years as interim, and seven to ten years as long-term outcomes. These time frames should be adjusted to fit the specific intervention you are implementing and evaluating.
may influence both the implementation of a program and the subsequent outcomes of that program and explain why those outcomes differ across sites. For example, school leaders may belong to professional unions with varying levels of power, work in cities with different economic or political conditions, or operate in schools or districts that are also instituting other educational programs that may also influence the outcomes the school leadership intervention intends to address. Knowing contextual factors can help implementers proactively plan to address potential barriers to implementation. These contextual factors do not need to be identified for each individual school, but the broad themes (e.g., union support, political and economic context, competing or supporting programs) can be included for consideration in the program evaluation.

For example, a new state law may require school leaders to receive annual evaluations tied to professional goals. The specifics of the law could affect the district’s leader evaluation system if, for example, the law required student achievement to be a factor in evaluations.

The components of the logic model will be the same across program types. However, the specific elements within each component may differ by program type. For example, principal preparation program activities will likely involve coursework and internships, while working conditions activities will likely have to do with district policy in such areas as principal pay and autonomy.

**Assembling Your Logic Model**

Tool 2.1 provides a structure for organizing the logic model information and drawing the connections between each of the key components. To assemble the intervention-specific logic model, first identify the important inputs, activities, outputs, and outcome related to your intervention. Chapter 4 provides some thoughts on outcomes, and information from the intervention developer—as well as conversations with your colleagues and the local implementers—can help identify inputs, activities, and anticipated outcomes. You may also reference the *School Leadership Interventions* report for additional ideas. Another companion piece to the *School Leadership Interventions* report is *Logic Models for Selecting, Designing, and Implementing Evidence-Based Leadership Interventions* (Daugherty, Herman, and Unlu, 2017). The professional learning logic model from this companion piece, below (Figure 2.1), demonstrates how a tailored logic model might look. Daugherty and colleagues (2017) provide detailed, completed logic models for each of the six types of leadership activities and can assist in the identification of resources, activities, outputs, and outcomes.

Place each input, activity, output, and outcome into its own box (with more boxes being added where needed). Using this basic structure, you add, subtract, or move arrows connecting boxes to identify the interrelationship between the components of the logic model. The connections should follow resources through to long-term outcomes, clearly articulating how resources are expected to produce the identified outcomes, through activities and outputs.
Tool 2.1: Logic Model Template

Program Name:  
Date of Logic Model:  

Resources  Activities  Outputs  Short-Term Outcomes  Intermediate Outcomes  Long-Term Outcomes

Contextual Factors
Logic models can be dynamic. As the team developing the logic model challenges assumptions or discusses how the program functions in a specific context, more information can be added, or connections between components can be altered. Additionally, the program logic may change over time when resources, activities, or contextual factors also change. Dating a logic model will help administrations and the evaluator track these changes throughout the life of the program.

**Figure 2.1: Logic Model, Professional Learning**

- **Resources**
  - Human resources
  - Facilities
  - Funding
  - Materials

- **Activities**
  - Select or develop evidence-based professional development curriculum
  - Identify and train (as needed) professional development providers (trainers and coaches) on the curriculum/program structure
  - Align professional learning with principal needs
  - Provide principals with high-quality training (i.e., evidence-based, applied, intensive) that contributes to career progression
  - Provide principals with access to regular, just-in-time support
  - Develop peer learning communities and facilitate engagement

- **Outputs**
  - Program content and structure is aligned with evidence on effective leadership
  - High-quality trainers/coaches/mentors prepared to deliver training to principals
  - Principals receive professional learning aligned with their needs
  - Principals receive sufficient, high-quality support during formal training sessions
  - Principals receive sufficient, high-quality support between formal training sessions

- **Outcomes**
  - Improved principal competencies
  - Improved schools
  - Improved student achievement

- **Indicators**
  - All reviewed programs undertook this activity
  - Some reviewed programs undertook this activity
  - All reviewed programs mentioned this output
  - Some reviewed programs mentioned this output
Having a complete and accurate representation of the program logic is crucial for the evaluation of programs. The team should agree that all critical elements are present. The components of the logic model will shape the type of information collected for the evaluation, how it is collected, and the data analysis. Tool 2.2 provides a few questions that teams can ask to verify the accuracy of the completed logic model (adapted from McLaughlin and Jordan, 2010, p. 72).

**Tool 2.2: Evaluating the Logic Model**

<table>
<thead>
<tr>
<th>Questions About the Logic Model</th>
<th>Responses</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Is the information in the logic model sufficiently detailed to identify the components and their interrelationships?</td>
<td>□ Yes    □ No</td>
</tr>
<tr>
<td>2. Is the program logic complete? Are all important elements accounted for in the model?</td>
<td>□ Yes    □ No</td>
</tr>
<tr>
<td>3. Is the program logic sound? That is, do all components fit together logically? Does the model account for all plausible pathways to achieving the program outcomes accounted for in the model?</td>
<td>□ Yes    □ No</td>
</tr>
<tr>
<td>4. Have the most powerful and malleable external contextual factors been identified and their influence on resources, activities, outputs, and outcomes identified?</td>
<td>□ Yes    □ No</td>
</tr>
<tr>
<td>5. Are the identified outcomes SMART (specific, measurable, action-oriented (or assignable), realistic, and timely)?</td>
<td>□ Yes    □ No</td>
</tr>
</tbody>
</table>

**Additional Guidance**

Should your team want more guidance for developing a logic model, understanding the critical components of a model, or reviewing examples of education-specific program logic models, here are a few recommended resources to assist you:

Summary
You should now be able to identify
✓ resources, activities, and outputs linked to your program
✓ whether outcomes are short term, intermediate, or long term
✓ local and external contexts that may affect the implementation or the outcomes of your program.

In the next chapter, you will learn how to assess potential evaluation approaches that are appropriate given the outcomes being measured, as well as the resources, expertise, and time available for the evaluation.
Chapter 3: Understanding Program Evaluation Methods

In this chapter, we provide a brief summary of the types of evaluation methods likely to meet the ESSA standards for evidence-based practices, and we describe how to select an evaluation method that will best demonstrate the outcomes associated with your school leadership program, based on the type of program you are implementing and your available resources. We also highlight factors that may affect your evaluation design. Importantly, this toolkit focuses on outcome evaluations only for school leadership interventions. The regulations in ESSA do not require state or local administrators to consider or develop an evidence base on program implementation. However, we recognize the importance of evaluating program implementation in order to (1) understand the factors that may or may not influence a program’s ability to produce intended outcomes, (2) identify ways to continuously improve ongoing programs, and (3) be an informative resource for other adopters of the same or similar programs. With this in mind, Appendix B provides a brief introduction to implementation evaluations, and we raise examples of how implementation evaluation measures can be examined with outcome evaluations.

Types of Outcome/Impact Evaluation

Outcome evaluations focus on the short-term, intermediate, and long-term outcomes expected as a result of implementing a program. Outcome evaluations address such questions as:

1. What effect did the intervention have on student achievement (long-term, student-level outcome)?
2. What effect did the intervention have on improving teachers’ instructional practices (intermediate-term, teacher-level outcome)?
3. What effect did the intervention have on reducing turnover of effective teachers (intermediate, school-level outcome)?
4. What effect did the intervention have on staff climate (intermediate, school-level outcome)?
5. What effect did the intervention have on principals’ guidance to teachers (short-term, principal-level outcome)?

We focus on three types of evaluation methods that address outcome-related questions. This is not an exhaustive list of analytic methods, but the highlighted methods align with the federally determined standards for evidence-based practices.

Experimental Studies

An experimental study provides evidence on whether the intervention—and not some unrelated factor, such as participants’ background—caused the outcomes. Because the study design is rigorous, findings are likely to be accurate and unbiased. According to the Education Department General Administrative Regulations (EDGAR), randomized control trials (RCTs), regression discontinuity design studies (RDDs), and single-case design studies (SCDs) are eligible types of experimental design research. The U.S. Department of Education’s What Works Clearinghouse (WWC) indicates that these study designs should meet certain criteria (such as low attrition) to classify as experimental studies (Tier I); if they do not meet those criteria, they are considered weaker studies and relabeled “non-experimental” (i.e., Tier II or III) evidence instead. The “Well-Designed and Well-Implemented Studies information” box later in this chapter provides more guidance on the challenges that are common in experimental studies. Guidance specific to RDDs and SCDs is available online at https://ies.ed.gov/ncee/wwc/Handbooks.

---

6 As described in Herman et al. (2016), the ESSA legislation defines evidence tiers but users may differ in how to apply the tiers. The approach used in the review, which informs this tool kit, is consistent with prior legislative and U.S. Department of Education approaches to evaluation program evidence of impact. See Herman et al. for more information.
A key feature for all types of experimental studies is that only the intervention group receives the intervention. In RCTs, some participants are randomly assigned not to receive the intervention (i.e., a nonintervention control group), and, in RDDs, some participants (i.e., a nonintervention comparison group) just miss assignment to the intervention group based on a specific cutoff (e.g., an evaluation score). Any difference in outcomes between the intervention and nonintervention groups can be attributed to the intervention itself. SCDs do not have a nonintervention group but include “phases” or periods in which the participants do and do not receive the intervention. For SCDs, differences in outcomes for participating individuals across phases (e.g., between preintervention and mid- or postintervention) are attributed to the intervention itself. Importantly, because of the explicit assignment to either intervention or nonintervention groups or identification of study phases in the case of SCDs, an experimental study requires advance planning and cannot be conducted once a program is already being implemented.

**Quasi-Experimental Designs (QEDs)**

Using a QED is the next-best alternative to an experimental study in terms of drawing causal inferences from the results of the data analysis. Common QEDs include propensity score matching and difference-in-differences estimations. These methods attempt to identify a comparison group that matches the intervention group prior to the intervention. Although QEDs often put participants into groups (nonrandomly) before the intervention starts, QEDs can be estimated using historical data. For the purposes of ESSA, QED research has the potential to produce Tier II or Tier III evidence, depending on how well the study was conducted.

**Correlational Analyses with Controls**

Studies using correlational analysis can identify the relationship between an individual’s participation in an intervention and the outcomes of interest. These analyses do not provide causal estimates but provide evidence that an intervention *might* be beneficial. To meet the Tier III evidence standards set in ESSA, correlational analyses must account for participant characteristics (e.g., years of experience, gender, languages spoken, prior measures of leader effectiveness) that may be related to the outcomes being measured. Regression modeling and interrupted time-series designs (which look at patterns of data at points before and after an “interruption” by an intervention) are two examples of correlational analyses that can control for prior characteristics. Both analyses can use historical data if preintervention information is available for participants and/or nonparticipants.

As suggested earlier, these three categories of evaluation methods map, respectively, onto the evidence tiers identified in ESSA legislation and further explained in the U.S. Department of Education guidance (U.S. Department of Education, 2016). Well-designed and well-conducted experimental studies have the potential to provide Tier I evidence; QEDs can provide Tier II evidence, and correlation analyses with controls produce Tier III evidence. We do not present detailed information on Tier IV evidence. However, Tier IV evidence requires a well-defined logic model, based on prior evidence, that aligns with regulations in EDGAR (see https://www2.ed.gov/policy/fund/reg/edgarReg/edgar.html for more information).

**Choosing Among Evaluation Designs**

Each of the foregoing evaluation methods offers benefits and challenges, detailed in Tool 3.1. The first column of the tool identifies the evaluation design (e.g., experimental designs or QEDs), followed by columns for the advantages and disadvantages associated with that particular design. The last column suggests the potential of the specific evaluation design for each of the six program types discussed at the end of Chapter 1. We made the determination of “fit” of evaluation design to program type based on (1) our judgment of the feasibility of the study design for that type of intervention, and (2) whether there are existing studies of that type of intervention using that study design. For example, most principal preparation programs involve candidates applying for the program based on such factors as proximity. Further, candidates pay tuition for most principal preparation programs. Under these conditions, it is hard
to imagine a situation where candidates would accept being randomly assigned into a preparation program. Further, there are no experimental studies of principal preparation programs. Therefore, we consider the experimental study design to be a poor fit for principal preparation programs.

**Tool 3.1. Fit of Evaluation Design to Program Type**

<table>
<thead>
<tr>
<th>Evaluation Design</th>
<th>Advantages</th>
<th>Disadvantages</th>
<th>Fit with School Leadership Improvement Interventions</th>
</tr>
</thead>
</table>
| Experimental study| o Most accurate estimation of impacts | o Participants may not be willing to be randomly assigned.  
  o Ethical issues may arise if programs or services are knowingly withheld from potential beneficiaries.  
  o Costs for recruitment and assignment may be higher than in other designs. | ?  X  X  ✓  ?  ? |
| QED               | o May use existing data, which can reduce effort and cost | o Potential for contamination of samples  
  o Can be hard to identify matched group for analyses  
  o Difficulty accessing existing data | ?  ?  X  ✓  ✓  ✓  ✓ |
| Correlational Analysis with Controls | o Uses existing data, which can reduce effort and cost | o Not rigorous impact evaluation  
  o Difficulty accessing existing data | ✓  ✓  ✓  ✓  ✓  ✓ |

Key: ✓ = good fit; ? = may be good fit for some interventions; X = bad fit

The experimental study designs produce the most accurate impact estimates (see, for example, Cottingham, Maynard, and Stagner, 2005). However, recruitment for an RCT can be challenging, as many potential participants are not enthusiastic about the possibility of being randomly assigned not to receive the intervention (Mosteller and Boruch, 2002). State and local administrators may feel that withholding
services from potential beneficiaries would be unethical. SEAs or LEAs can consider using a waitlist design approach, where the matched comparison or control group receives the intervention following the evaluation period to reduce any concern of withholding potential benefits from recipients over time. Because of challenges in finding willing participants, costs for recruitment may be higher than for other designs. Random assignment of participants may be impractical for some types of interventions. Despite these challenges, some districts can and do use experimental designs—the quality of the findings can balance the logistical and political challenges. Regression discontinuity is another type of experimental design. Regression discontinuity requires a cut point, with those below the cut (those most in need) participating in the intervention and those above not participating. This design is more politically viable, as the neediest are getting the intervention. The design requires a measure to use for the cut point, such as principal evaluation scores, to qualify for a professional learning intervention and multiple data points before and after the cut point.

- **Leadership evaluation system**: If the leader evaluation system is determined at the state level, it might be possible to roll out a new system in stages, randomly assigning one group of districts to the new system. If the evaluation system is determined at the district level, random assignment becomes harder—implementers are likely to try the new (or old) system with all their schools, regardless of assignment. Additionally, the sample size of schools within some districts may be too low for a quantitative study. Regression discontinuity might be a possibility, if—for example—principals scoring low on prior evaluations were assigned to an evaluation system designed to help them improve.

- **Principal preparation program**: It would be difficult to randomly assign principal candidates to different preparation programs because principal candidates have strong preferences (e.g., they may choose a program close to where they live or one affiliated with their district), and preparation programs may not have enough qualified candidates to allow a group to be assigned to a control condition. Regression discontinuity would also be difficult for the same reasons, and an additional reason: It is not clear what outcome measure would have multiple data points before and after the intervention.

- **Strategic staff management**: Because staffing within a district is often a closed system in which principals removed from one position are moved to another in the district, and because those decisions are generally made individually, based on principal qualifications and fit, it is difficult to imagine a case where the district could assign principals to work in or leave a school using random or regression discontinuity methods.

- **Professional learning**: Professional learning is often individualized, and, under many conditions, principals can be randomly assigned to participate in a professional learning intervention or assigned based on their needs, according to some premeasured variable.

- **Working conditions**: There have been well-conducted studies in which participants have been randomly assigned to working conditions (e.g., incentives). Some working conditions (e.g., school-based autonomy) may be more difficult to randomize than others.

- **School reform models**: It is possible to randomly assign schools to reform models, but it is difficult. Prior research suggests that staff commitment to a model contributed to the quality of implementation, and staff commitment is not assured when schools are randomly assigned. However, several school reform models have been evaluated with experimental designs, especially when demand for the model exceeds the provider’s ability to deliver. In those cases, some who want to implement the model may have to wait until after the study.

The second most rigorous study design, the QED, has demonstrated bias that can be partially, but not fully, met by well-matched comparison groups and controls for preintervention measures (Glazerman, Levy, and Myers, 2002). Several design features of the intervention or its context—such as a lack of existing data or a lack of a comparison group—may make a QED infeasible. Despite these potential limitations, the QED has important practical advantages over an experimental study. Using a QED eases recruitment challenges posed by an experimental study. Some QED evaluations may take advantage of existing data, rather than collecting all the data themselves. This can reduce the effort and cost of the evaluation.
• **Leadership evaluation system:** If the evaluation system is determined at the state level, it might be possible to roll out a new system in stages, nonrandomly assigning districts to different conditions. If the evaluation system is determined at the district level, a QED becomes harder, though not infeasible; the challenge is that districts are likely to try the new (or old) system in all their schools, regardless of assignment. Additionally, the number of schools within some districts may be too few for a quantitative study.

• **Principal preparation program:** Identifying a comparable group of principal candidates may be difficult, because the factors that cause a candidate to choose a program may make him or her fundamentally different from comparison candidates. For example, candidates who prefer to go to large state universities may differ from those who prefer alternate routes. However, it is possible to do a QED study comparing earlier and later cohorts within a school, or to statistically control for many of the differences among schools and principal candidates.

• **Strategic staff management:** Because staffing within a district is often a closed system in which principals removed from one position are moved to another in the district, it is difficult to imagine a case where moving principals to or from schools could be adequately separated, in the analysis, from other conditions that might affect outcomes.

• **Professional learning:** Professional learning is often individualized, and under many conditions principals can be assigned to a group that will participate in a professional learning intervention or a group that will not participate.

• **Working conditions:** Some interventions involving working conditions, such as providing incentives to principals for improving student outcomes, may be suitable for QEDs.

• **School reform models:** QEDs can be and have been used to evaluate school reform models. For example, a large district may offer low-performing schools the choice of two reform models and compare the results. Because so many factors may differ between the schools (e.g., student demographics, teacher quality), isolating the effects of the intervention from other factors may be especially difficult.

The least rigorous of the Tiers I–III evaluation designs, correlational analysis, is also the easiest to carry out. Evaluators use existing data rather than collecting new data. State or district administrative sets are relatively accessible, especially if the state or district is involved in implementing the intervention to be studied. Using existing data can reduce the effort and cost of conducting the evaluation. However, analysis is limited to the data already collected, which might not match the questions asked in the evaluation. Further, the findings are exploratory only, because this design introduces substantial potential bias. All types of school leadership interventions may be evaluated with correlational analysis, with the caveat that the findings are not conclusive, given the limited rigor of the design.

**Quantitative Versus Qualitative.** Studies involving quantitative analysis (with valid and reliable measures), when well conducted, typically involve a larger sample of participants than qualitative studies and may be more generalizable to other educational contexts. However, we note two challenges related to quantitative studies. First, these studies require large sample sizes to identify effects (typically 50 or more)—smaller studies may miss real effects. Because principal interventions typically happen at the school level, this means the study will need to look at a large number of principals or schools. That is not always possible. Second, studies designed to capture the quantitative impact of an intervention on specific measures may not measure (and so not report) unintended consequences (Bamberger, Tarsilla, and Hess-Biber, 2016). A mixed-method study, which collects quantitative data on prespecified measures and exploratory qualitative data throughout implementation, is more likely to capture effects that were not necessarily part of the original logic model and effects that cannot be measured quantitatively. Thus, a well-conducted quantitative study may provide stronger information on impacts, and a well-conducted qualitative study may provide richer detail on implementation (see Appendix B).
**Tip: Well-Designed and Well-Implemented Studies**

In addition to using an appropriate method for measuring the outcomes of an intervention, it is important for evaluators to understand what makes a well-designed and well-implemented evaluation. The WWC has established standards for high-quality research, or how to determine if an evaluation study is well designed. The current guidelines are available at [http://ies.ed.gov/ncee/wwc/Handbooks](http://ies.ed.gov/ncee/wwc/Handbooks). The WWC requires studies to use either experimental or QED evaluation methods and meet additional criteria. The overarching goal of these criteria is to ensure the credibility of the evidence produced by the evaluation. In other words, the effectiveness of the intervention identified by the evaluators is likely to be the actual effectiveness of the program. The four criteria to consider specific to RCTs and QEDs are sample attrition, baseline equivalence, outcome eligibility and reporting, and confound.

1. **Sample attrition** sets limits for the proportion of the original evaluation sample (intervention and control participants) that has missing outcomes data. There are guidelines for the attrition rates of the overall sample, as well as for the difference in attrition rates between intervention and control groups (called differential attrition). As overall attrition increases, differential attrition allowances decrease. Exact allowance rates can be found in the current WWC Procedures and Standards Handbook.

2. When sample attrition is high or a QED method is used, the evaluators must demonstrate **baseline equivalence**: that the comparison group is not statistically different from the intervention group prior to the intervention (at baseline). Some analytic adjustments are made when differences are within a preestablished range of difference (see the Procedures and Standards Handbook).

3. The **outcome eligibility and reporting** must (1) meet face validity and reliability standards, (2) not be overly aligned with the intervention, and (3) be collected in the same way for both intervention and comparison groups. In other words, the assessment should measure what it is supposed to measure (e.g., reading tests should assess reading competencies), the assessment should produce similar results under similar assessment conditions (e.g., time limits, external disruptions), and the assessment should not use materials from the intervention itself.

4. **Confound** is the concept that no condition would have affected one group but not the other, with the exception of the intervention itself. For example, if the school leaders in the intervention group all happened to work in schools that received additional award funding for the year and the control group principals did not work in schools receiving this funding, the findings might be confounded. The effect of the intervention could not be distinguished from the effect of the funding on the measured outcome.

---

**Identify Issues Pertinent to Your Evaluation Design**

Several factors influence program evaluation design. Acosta and colleagues (2013) identified five factors that evaluators should keep in mind when selecting and designing their evaluation. We identified two other factors (availability of comparison groups and school leader placement within them) that are particularly relevant to school leadership programs. We define these factors and explain how they relate to evaluation design. Readers interested in learning more about the other factors are encouraged to read the Acosta et al. report available at [http://www.rand.org/pubs/tools/TL111.html](http://www.rand.org/pubs/tools/TL111.html).

1. Number of program participants
2. Links between principals and students
3. Timing of the evaluation
4. Data security and human-subjects considerations
5. Evaluation expertise
6. Resources available for the evaluation
7. Availability of comparison groups.
Number of Program Participants

The number of participants in a program likely shapes the information an evaluator can garner about that program. Statistical analyses require a minimum sample size—or number of participants and comparison group individuals—to identify the impact a program has on the intended or desired outcomes. Without a sufficient number of participants on whom to base an evaluation, even the best-implemented programs will fail to generate valid and reliable evidence of their effectiveness. According to EDGAR, the research base for an intervention should be based on more than one site and should include at least 350 participants; this might be achieved by combining multiple studies.

Factors that can shape sample size requirements include the types of outcome measures to be analyzed (e.g., student test scores or the evaluation scores of newly hired teachers); the effects of similar programs on similar outcomes measures (these provide a benchmark or guide for evaluators); and the resources available for both implementing and evaluating the program. While large sample sizes are commonly preferred, states may not be able to serve hundreds of participants with every program. Working with a statistician prior to implementing a program can help determine the appropriate balance between program resources and evaluation needs.

Links Between Principals and Students

In addition to having enough school leaders participating in the study for statistical analysis, evaluators should also link data about school leaders to data about students, using student-level data if possible. Studies that do not use student-level data, such as studies that analyze the relationship between principals and school average achievement, produce less-powerful and less-accurate analysis. First, the analysis is less powerful because there are fewer cases (say, ten schools instead of 20,000 students). Using students (when statistically adjusted for the students being in the same school) amplifies the power of the analysis and improves the likelihood that the study will find a connection between intervention and outcomes if a connection exists. Second, students move in and out of schools between and during the school year. Each year, a new class of students enters and another class exits the school. In addition, especially in low-performing schools, students transfer in and out of schools mid-year for various reasons, such as a change in housing arrangements or seeking private or specialized schools. Linking school leaders in the study to individual- rather than school-level outcomes increases the accuracy of the analysis by focusing on students who actually attended the school when the program participant worked there.

Timing of Evaluation

Two types of timing inform the evaluation. First, the length of the program will influence when outcomes can be assessed, when impacts are expected to be realized, and, potentially, the size of the impact. Second, the time available for an evaluation may also determine which outcomes (i.e., short term, intermediate, or long term) can be measured, or the types of data and resources the team will be able to collect and analyze. Interventions designed to improve school leaders may have shorter-term impact on teachers but may take some time to effect changes on student outcomes. Understanding program and evaluation time lines can help evaluators select appropriate and feasible designs.

Data Security and Human-Subjects Considerations

Evaluators must ensure the safety and privacy, where appropriate, of research participants and the individuals who are part of the program evaluation. Administrative data sets, including personnel data or student test-score files, typically require well-established security protocols and specific regulations for who can access and utilize the data for evaluation purposes. This information must be protected against inappropriate access and use.
Research on sensitive topics, such as principal performance ratings, or that includes young children requires additional regulations be instituted prior to beginning an evaluation, to protect the individual’s wellbeing. The U.S. Department of Education provides information regarding human-subjects protection at: http://www2.ed.gov/about/offices/list/ocfo/humansub.html. State offices of education or local education entities may have additional documents or regulations (e.g., parental notification) that must be accounted for prior to conducting the evaluation.

**Evaluation Expertise**

The type of evaluation selected for a particular program will depend on the expertise of those individuals responsible for the evaluation process. If evaluators are trained in rigorous analytic methods, a more-rigorous design may be possible. If the evaluator does not have advanced statistical training, conducting an experimental design or implementing QEDs may not be feasible. In selecting an evaluation design, SEA and LEA administrators should consider the technical capacity of their staff and/or the resources available to contract an external evaluator that has the needed skill set. Consulting with academic institutions or a research organization prior to implementing a program is recommended when contracting these external partners. The American Evaluation Association (AEA) provides a searchable database (http://www.eval.org/find_an_evaluator/evaluator_search.asp) of members available for evaluation consulting. Similarly, the WWC Help Desk (https://ies.ed.gov/ncee/wwc/Help) provides support to evaluators aiming to develop rigorous studies.

**Resources Available for the Evaluation**

All program evaluation requires some expenditure of SEA and LEA resources, such as staff time, data storage or computer technology (e.g., statistical software), and financial resources (e.g., paying for external evaluators). The available resources will inform the feasibility of particular evaluation designs.

Identifying the available resources prior to program implementation can help minimize resource burden and maximize the benefit gained from the resources expended. For example, taking an inventory of available data sources (e.g., surveys, administrative data, and observation rubrics already collected as part of the program) can help reduce duplication of efforts, saving both time and money. The program logic model can help evaluators identify the types of data and resources needed to evaluate the program (e.g., measures for analyzing short-term outcomes).

**Availability of Comparison Groups**

High-quality program evaluation strives to compare outcomes for a school leader “treated” by a program with outcomes for a school leader who was not treated but is otherwise as similar as possible to the “treated” school leader. Finding good comparison groups for school leadership programs is difficult. Not only do school leaders vary in terms of their individual characteristics (e.g., age, race, years of experience), but the schools and districts in which they are placed also vary. These differences in school and district characteristics matter for the evaluation. For interventions targeting current school leaders, random assignment to the program for some school leaders but not others may be feasible. However, if the program is intended to support leaders in particular types of school (e.g., turnaround schools), comparison groups may be limited. Even if there are sufficient similar schools to support a randomization, it may be difficult for a district to withhold a program from some schools while offering it to others. It is important to think through the effective construction of comparison groups when designing an evaluation and build support for randomization where it is feasible.

**Additional Resources**

For more information about outcome evaluations and the identification of appropriate evaluation methods, here are a few recommended resources to assist you:


**Summary**

You should now have a clear understanding of

- the different types of outcome evaluations
- how evaluation designs map onto ESSA evidence tiers
- key factors that will affect your evaluation design.

In the next chapter, you will learn how to assess and select outcome measures for your evaluation.
Chapter 4: Identifying Outcome Measures

In this chapter, we focus on the types of outcome measures frequently utilized in school leadership intervention evaluations. We review the types of data that you may have available and the advantages and disadvantages associated with each.

Tip: Overcoming Challenges with Measuring School Leadership Interventions

In the first chapter of this toolkit, we highlighted five challenges associated with principal program evaluation. Two of these challenges related to the outcome measures commonly utilized in evaluations: First, it may take multiple years before impacts on student outcomes can be identified; and second, focusing solely on student achievement outcomes does not capture the full scope of impact from interventions. Evaluators should consider measuring short-term and intermediate outcomes (a) when time does not permit long-term outcome evaluations, and (b) to identify the extent to which earlier outcomes are or are not met. Measuring the impact of the program on both short-term and intermediate outcomes will also assist in interpreting and understanding long-term effects once they are identified.

Short-Term, Intermediate, and Long-Term Outcomes

In the discussion of logic models (Chapter 2), we described short-term, intermediate, and long-term outcomes. For school leadership interventions, short-term outcomes could be changes in the principals, intermediate outcomes could be changes that the principal directly affects (i.e., teachers’ instruction, school climate), and long-term outcomes could be student-level improvements.

Guskey’s (2014) five levels of teacher professional development evaluation provide additional guidance on thinking about short-term outcomes for interventions that involve learning (e.g., principal preparation, professional learning, leadership evaluation systems). Applying Guskey’s levels to principals, short-term outcomes might include:

- Principals’ learning of new knowledge and skills: This information might be collected with a test or performance assessment.
- Principals’ use of the new knowledge and skills: This information should be collected after the program. The strongest measures include direct observation or performance measures. Although some evaluations use principals’ self-reports of their use of the knowledge or skills, that measure is prone to bias.

For principal interventions that involve changing context (e.g., working conditions and some elements of leadership evaluation systems), short-term outcomes might involve using behaviors consistent with the new policy (e.g., using new authorities under working conditions or implementing a new curriculum under school reform).

Intermediate outcomes indicate that the changes made by the principal have affected the school in important ways. If it is not reasonable to measure long-term, student-level outcomes in your evaluation, intermediate outcomes are quite critical. These outcomes should focus on class- or school-level changes that have been found in prior research to affect important student outcomes (as you described in your logic model). For example, intermediate-level outcomes might include increased time on task in classrooms across the school, fewer disciplinary incidents, or improved student and teacher attendance. The specific intermediate outcomes depend on the content focus of the intervention: professional development on mentoring teachers may affect instructional quality, and professional development on a behavior management program may affect disciplinary incidents.
Long-term outcomes typically include student achievement but also may include increased student success in terms of student high school graduation rates, college or career outcomes, or students’ social and emotional wellbeing. For example, a principal-focused intervention designed to help improve student academic performance in high school, such as school turnaround training, would be expected to measure student achievement and perhaps high school graduation and college entry rates. While short-term and intermediate outcomes may differ by type of leadership intervention or specific intervention, long-term outcomes are typically the same regardless of intervention: improvements in student-level outcomes, especially achievement.

Level of Outcome Measure

Outcome evaluations can address multiple levels of performance, given the right types of data. For a school leadership intervention, studies might be conducted on the intervention as a whole. Such an evaluation would look at the average impact an intervention demonstrates across all sites and participants. An intervention might be measured at a site level, meaning how a program functions in each city, district, university, or location that delivers or receives the intervention. Program outcomes might be best analyzed at an individual principal or participant level. The outcomes identified in the logic model and the overall goals for the outcome evaluation, in conjunction with the available data sources, will shape the level at which the analysis is conducted.

Utilizing Measures Used in Prior Evaluation Studies

As a good start to selecting measures for your evaluation, you can identify studies that measure the same or similar outcomes as those in your program’s logic model. If exact outcome matches are not found, consider whether one or more of the previously used outcomes is a strong proxy for your outcome of interest. Here are some commonly used measures:

- How principals spend their time: coaching, supervision, instructional activities, interacting with families
- Principal evaluation scores: supervisor evaluations, 360 evaluations (e.g., Val-Ed)
- School conditions: school climate, parent involvement, recruitment and hiring of teachers (e.g., number of applicants, qualifications of applicants), readiness of newly hired teachers, effectiveness of teachers, teacher retention, staff climate indicators
- Student outcomes: achievement on standardized tests; measures of academic growth; rates of disciplinary actions, attendance, or dropout; student course-taking or rates of college- and career-readiness; student climate and engagement.

In Appendix A, we detail studies on school leadership that have ESSA Tier I–III evidence. The outcomes of those studies and the specific outcome measures used are identified in Table 4.1. We encourage program evaluators to consider the measures identified in that table a starting point rather than a comprehensive list. In fact, two program types do not have readily identifiable measures from existing literature. Evaluators are encouraged to explore measures beyond this list to ensure that the most appropriate measure is selected for the given program’s context and population, as well as for the study question guiding the evaluation.
Tool 4.1. Outcome Measures Used, by School Leadership Program Type

<table>
<thead>
<tr>
<th>Program Type</th>
<th>Short-Term Outcome (Measure)</th>
<th>Intermediate Outcome (Measure)</th>
<th>Long-Term Outcome (Measure)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal preparation</td>
<td>• Principals’ working conditions (self-report survey)</td>
<td>• Quality of teaching staff (alignment between certification status and subjects taught)</td>
<td>• Student achievement (standardized test scores, school proficiency rate)</td>
</tr>
<tr>
<td></td>
<td>• Principals’ use of time (self-report survey)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Principals’ retention (years retained in same school)</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>• Leadership behaviors (self-reports and ratings from peers on evaluation survey)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Professional learning</td>
<td>• Principal leadership practices (McREL-aligned survey of leadership qualities)</td>
<td>• Teacher and principal turnover (retained in same school for 3 years)</td>
<td>• Student achievement (standardized test scores, school proficiency rate)</td>
</tr>
<tr>
<td>Working conditions</td>
<td></td>
<td></td>
<td>• Student achievement (standardized test scores, school proficiency rate)</td>
</tr>
<tr>
<td>Leadership-focused school reform models</td>
<td>• School attendance rate (average daily attendance)</td>
<td>• Student achievement (standardized test scores, school proficiency rate, school performance rating)</td>
<td>• Student achievement gains (change in standardized test score in subsequent years)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>• Student motivation and engagement, behavior, satisfaction (parent and student self-report surveys)</td>
</tr>
</tbody>
</table>

Identifying and Selecting Sources of Data

You should consider two types of data sources when designing your evaluation: accessing data from existing sources and collecting new data as part of the evaluation. There are advantages and disadvantages associated with both categories of data, summarized in Tool 4.2. You will likely use a combination of sources to meet all of your evaluation needs. Up-front planning can help you make the most effective use of existing data, design data collection to serve multiple purposes, and reduce the burden of data collection on program participants.

First, consider using existing data, which were collected for purposes unrelated to the evaluation. These data include the administrative data collected each year by SEAs, LEAs, and schools on their students, staff, and outcomes, as well as program-specific data collected by school leadership programs to guide implementation. Existing administrative data can be particularly useful for QED evaluations (Tier II) and for selecting a sample for an experimental study (Tier I). Examples of existing administrative data include student achievement data, attendance data, data on school personnel from human resources.
systems (e.g., retention, turnover, evaluation scores), data on administrative certifications, etc. Examples of program-specific existing data include program curricula; participant application, enrollment, and attendance records; and interim assessments or feedback on program participants.

Remember—these data are not collected for your evaluation. You should take care to understand the limitations of the existing data and the reasons the data are collected (e.g., compliance with federal regulations, determining program funding levels). These data collections may change over time, and evaluators should track those changes as they may influence conclusions of an evaluation. There are several drawbacks to using existing data: (1) Data collected to meet other requirements may not exactly fit the needs of the evaluation, and (2) program-specific data are not available for the comparison group. Notably, administrative data collections guided by state or federal regulations likely cannot be tailored to evaluation needs; however, program-specific data collections may be effectively tailored to meet both operational and evaluation needs when evaluators work with program staff.

If you plan to use existing data for an evaluation, check for the following:

- **Documentation about the data**: Programs should document the decision rules or rubrics used to assess participants and document changes to those rubrics over time. If you “tweak” a certain measure to better align with your theory of action, make a note of how the data changed. Documentation is important for the evaluation—especially if the people most knowledgeable about the data leave your SEA or LEA.
- **Year-to-year retention of data**: If the evaluation requires multiple data points (say, attendance rates for the last three years), make sure the data are retained from one year to the next.
- **Data accuracy**: In any existing database there may be mistakes or errors in the data that are inconsequential for the original purpose but critical for your evaluation. Build in checks and validations to make sure the data are correct.

**Evaluation data** are data collected for the purpose of the evaluation. These data can fill in shortcomings of existing data collections and are tailored to the program and population being evaluated. As mentioned previously, program staff and evaluators should collaborate prior to implementation, aligning the program and evaluation's data needs wherever possible. This helps ensure that data collection efforts do not put too much of a burden on the program staff or participants. In general, evaluation data have similar advantages and disadvantages as program data.

Regardless of data source, data sets may include private, personally identifiable information (PII) or sensitive information. When gathering or accessing PII or sensitive information, be sure to have robust informed consent and data safeguarding plans in place. Confidentiality and data use agreements are often required for the use of such information by Institutional review boards or other organizations designed to protect human subjects. Refer back to the information and links in the “Data Security and Human-Subjects Considerations” section in Chapter 3.
## Tool 4.2. Advantages and Disadvantages of Data Sources by Type

<table>
<thead>
<tr>
<th>Data Source Category</th>
<th>Examples of Types of Data Available</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
</table>
| **Existing data**    | - Personnel data (e.g., work experience, salaries)  
- Certification data (e.g., principal certification)  
- Workforce data (e.g., open principal positions)  
- Evaluation data (e.g., results of prior evaluations)  
- Student achievement data (e.g., scores on accountability tests)  
- Student enrollment, attendance, and discipline data  
- School accountability data (e.g., school ratings)  
- Program delivery or other operations data (e.g., number of sites, nature of professional development, participant acceptance rates)  
- Participant outcomes (e.g., self-reported improvement, evaluations) | - Evaluator does not bear the cost of data collection  
- Data may be available historically  
- Data typically are comprehensive, covering an entire population  
- Not burdensome for program participants or beneficiaries | - Evaluator does not influence what is collected  
- Evaluator does not influence when data are collected or recorded  
- Evaluator may misunderstand the data or be unaware of data limitations  
- Evaluator may incur a cost to access administrative data  
- [For program-specific data] Baseline data may not be available for the period preceding the intervention  
- [For program-specific data] Data may not be available for comparison groups |
| **Evaluation data**  | - Written assessments of school leaders’ skills or knowledge assessments  
- Surveys of students, teachers, other staff, and principals on principal performance, school climate, instruction, or similar topics  
- Direct assessments of teachers’ instruction or climate (e.g., through observations) | - Evaluator determines what is collected and when  
- Evaluator is aware of data limitations | - Evaluator bears the cost of data collection for intervention and control group  
- Data collection can be burdensome for the program and program participants and beneficiaries  
- Data on baseline practices are not typically available prior to the implementation of the program and may not be available for a period after the program begins |
Choosing Key Data Points (Indicators) for Your Evaluation Measures

To evaluate school leadership interventions, you may need access to data about different individuals and processes. First, you will need information about the individuals who participate in the intervention (usually aspiring or current principals). Second, you will need information about the program staff or individuals delivering the program. Third, you will need information about teachers and students in the school served by the program participant (ideally, past, present, and future).

Further, there may be specific relationships that you are interested in understanding. For example, maybe there is a state policy push to improve the effectiveness of early career principals. In this example, you will want to know how an intervention works for new principals (e.g., less than two years in the role), relative to mid-career (e.g., up to five years of experience) and experienced principals (e.g., more than five years). You may also be interested in comparing early career principals from one program with those early career principals who completed a different type of preparation program. Therefore, it is important to collect data about principal experience and perhaps structure the evaluation to include only early career principals.

Next, we present a more detailed list of indicators that you may wish to include. Note that these are examples and are neither a required nor comprehensive list of potential indicators.

**Information about program participants:** This will include basic demographic information as well as information about experiences, performance, and activities prior to, during, and after the intervention takes place. Relevant information may include career and educational experiences, performance in professional roles, scores on assessments, self-evaluations or reflections, and evaluations conducted by others.

**Prior experiences:** Examples include years and types of teaching experience, outcomes of students taught, years of administrative experience, certifications, and scores on assessments. Sources include administrative data, questionnaire or survey, and screening assessment administered by program.

**During program:** Examples include scores on assessments administered by the program, ratings by program staff, self-assessments by program participants, and satisfaction surveys by participants. Sources include program or evaluation data collected from participant or program staff (questionnaire, survey, assessment) and administrative data (if program is part of school district).

**After program:** Examples include scores on assessments administered by program, ratings by program staff, self-assessments by program participants, satisfaction surveys by participants, and placement information. Sources include data collected by program or evaluation staff from participant (questionnaire, survey, assessment), and administrative data.

**Information about program staff and consultants:** In addition to tracking information about program participants, it may also be useful to track information about program staff and consultants. This information could be used to explore relationships between program staff and variation in outcomes among participants. For example, if all of the participants coached by a particular coach have good outcomes, you may want to understand how that coach that was different from the others. Relevant information might include background characteristics, such as career history, education, training obtained, experience in role, and experience in the district. You can track the assignment of program staff to program participants and evaluations of program staff.

**Information about outcomes:** Finally, you will want to obtain information about students, staff, and other stakeholders in schools served by the program participants, such as school outcomes (school climate, student outcomes, teacher turnover, teacher satisfaction) or student outcomes. This information will likely be obtained from administrative sources, but you may want to conduct your own surveys or assessments to supplement administrative information. You should plan for any original data collection well in advance.
Determining Format for Evaluation Results

The format in which evaluation results are reported should be decided as part of selecting the measures. Some formats may be clearer than others for the intended audience, some may be needed for state or federal reporting requirements, and some may be more policy-relevant. Using the outcomes from the logic model, what information about a principal’s ability to identify and support effective instruction is useful? The following are a few examples:

- Frequencies or counts: Principals identify effective instruction 20 percent of the time or on 20 occasions. Prior to intervention, effective instruction was identified only 5 percent of the time or on five occasions.
- Overall proportions: 85 percent of participants demonstrated improved capacity to support effective instruction.
- Distributions of performance: 15 percent of participants showed no improvement, 30 percent showed moderate improvement, and 55 percent of participants demonstrated dramatic improvement.
- Comparisons of participants from separate program sites: Program participants in California demonstrated twice the capacity for identifying effective instruction as participants from Florida.
- Which participants demonstrated improvement: First-year principals improved three times as much from the training as experienced principals.

Did the Outcome Change “Enough”?  

Your logic model shows outcomes you hope to achieve through your school leadership intervention. Your analysis shows the outcomes you achieved. You can structure your outcomes and analysis to show whether your findings meet your expectations.

Statistical significance: The evaluation analysis will likely indicate whether the results are statistically significant. For example, if you are aiming to improve student achievement and findings appear to show gains, tests of statistical significance can signal that your findings are accurate and achievement did change. Typically, statistical significance at the 0.05 or 0.01 level is sufficient (Craparo, 2007; Sproull, 2002). Sometimes, however, tests of significance do not tell you enough. You may also want to know whether the gains are meaningful in the lives of students.

Effect size: Effect sizes provide information about the size of the impact. Effect sizes can range from very small to very large. Although interpreting effect sizes as small or large depends on the intervention, population, and context, some rough parameters are that 0.20 is generally a small effect size and 0.80 is generally a large effect size, using Cohen’s d to compute effect size (Cohen, 1988; Sawilowsky, 2009). For leadership interventions, in specific, there are no commonly acknowledged bars for “small” or “large” effect sizes. However, school leaders can account for approximately 25 percent of school-based changes in student achievement, and raising a school leader’s skills on a set of leadership qualities by one standard deviation can improve test scores by 10 points (Creemers and Rezigt, 1996; Waters, Marzano, and McNulty, 2003). Therefore, an intervention that improves a school leader’s skills can be expected to have measurable effects on student outcomes.

Benchmarks: In assessing whether an intervention accomplished the desired effect, it may help to set benchmarks. The final benchmark, of course, should be the goal. For example, one goal might be that 100 percent of students are reading on level by fifth grade. Another goal could be that (statistically) significantly more teachers report that their principals provided useful feedback. Benchmarks can be absolute (e.g., percentage of students who are reading on grade level) or relative (e.g., more teachers find feedback useful).
Interim benchmarks can help an evaluator determine whether adequate progress is being made. For example, 60 percent of students are reading on grade level in the first year of the intervention, 80 percent in the second year, and 100 percent in the third year. Or principals begin providing feedback with a new method in the first year, they revise the approach in the second year, and teachers find the feedback more useful by the third year.

Benchmarks should be set based on the goals of the intervention and what prior research suggests about school leadership interventions (see Appendix A) with adjustments for differences in the current intervention, intervention dosage, and who the program participants are in your intervention. These benchmarks should also be related to the pertinent outcomes, level of analyses, and specific relationships on which you have asked the analysts to focus. Because interventions, goals, and time lines differ, and because the research on benchmarks is limited and cannot be generalized to most situations, the examples in Appendix A are just that—examples.

If results do not meet goals, what would you want to know? How close to expectation did the intervention get? Whom did the intervention work for, or at what site was the intervention most successful at meeting these benchmarks? At what site were the intervention results furthest from desired outcomes? What happened at this site?

Analysts can also help describe the many ways in which the outcomes of interest can be measured and presented. The more explicit the analytic plan is prior to beginning data analysis, the more tailored the information can be for the evaluator, which means a more-effective use of limited resources.

**Summary**

You should now have a clear understanding of

- how to make use of outcome measures from previous evaluations
- what are the different categories of data available to measure
- the advantages and disadvantages of using these different data sources
- what data points you will likely want to collect for your evaluation.

In the next chapter, you will learn how to organize and analyze the data you have collected.
Chapter 5: Organizing and Analyzing Your Evaluation Data

In this chapter, we describe how to examine your collected data in a way that is accurate and useful by creating well-organized databases and provide resources to learn about basic analysis techniques. We also highlight some key factors to consider in working with your data analysis team.

Your approach to organizing evaluation data and performing analyses will vary depending on the type of evaluation you are conducting. If you are using individual-level data, you will want to be able to link information about program participants or beneficiaries over time. For example, you may want to link an aspiring principal’s preservice application score to assessments conducted during the program and placement information. You may also want to be able to link principals to their district-level supervisors, as well as to students in the schools the principals serve.

Organize Evaluation Data into Databases

Since data will come from multiple sources, you will need to think about how you will link information across those sources. In most administrative data systems, people, places, and programs all receive unique identifiers so that information from multiple data sources and from multiple years can be connected. Using these unique identifiers, a principal can be linked to the preservice training program she participated in, details about how much training she received while participating in that program, and her performance once placed into a school as a principal. This linkage is critical when trying to measure or compare the impact of school leadership interventions.

Analyze Evaluation Data

Once the evaluation data are assembled, you can analyze the data to address key questions suggested by the logic model:

- Was the program implemented as expected?
- Did program participants complete the program?
- Did program participants like the program?
- Did the knowledge, skills, and abilities of program participants improve?
- Did teacher practices and instruction improve?
- Did the school outcomes improve?

The more data that you have about the program participants and program staff, the more analysis you can do about the “whys” behind the answer to these questions. For example, if you learn from an analysis of program completion data that only 85 percent of the program participants complete the program, you might want to know why some people complete the program while others do not. You might examine whether program completion is related to the participant’s prior experience, the program staff providing the intervention, or baseline assessment scores.

There are many ways to analyze your evaluation data. We summarize three options here: (1) using Excel for basic analyses, (2) using a statistical software package, and (3) hiring an external evaluator to perform the analyses. If your internal team does not have an experienced analyst available for the analyses, we recommend seeking professional help.
Microsoft Excel can be used to conduct some basic descriptive analyses (e.g., summarize participant characteristics); analyze data from before, during and after the program to see whether program participants' knowledge, skills, behaviors, or other characteristics change; and determine whether program participants’ attendance, satisfaction, or other characteristics contributed to changes in their knowledge, skills, behaviors, or other characteristics.

**Tip: Using Excel to Analyze Data**

The *Suicide Prevention Program Evaluation Toolkit* provides three analysis primers that describe how to use Excel 2010 to conduct descriptive analyses, run statistical models for detecting differences in your program’s target population, and link process and outcome data. In the following table, we summarize the primers available in that document and the types of analyses described in each.

<table>
<thead>
<tr>
<th>Primer</th>
<th>Types of Analyses</th>
</tr>
</thead>
</table>
| Primer 1: Calculating descriptive statistics for your program | Describe the key characteristics of program participants  
Summarize attendance  
Describe participant satisfaction  
Calculate averages and other descriptive data (e.g., percentages, frequencies, ranges, modes) for each outcome variable |
| Primer 2: Statistical models for detecting differences in your program’s target population | Analyze pre- and posttest data to see whether program participants’ knowledge, skills, behaviors, and other characteristics change (e.g., does the program achieve its outcomes?) |
| Primer 3: Linking implementation and outcome measures | Determine whether program participants’ attendance, satisfaction, or other characteristics contributed to changes in their knowledge, skills, behaviors, and other factors. |

There are statistical software packages that help automate many analysis functions. For example, education researchers frequently use R, SAS, Stata, SPSS, HLM, and similar packages. Statistical packages afford several advantages, such as being able to conduct interrupted time-series analyses that are likely to be useful for programs that collect data at multiple time points. Statistical packages make it much easier to address known problems in the data, such as biases introduced by similarities of people within classes or schools. Users of these packages must have experience with the technical language of
the software, as well as statistical knowledge to navigate and apply the software capabilities to the administrative, program, and evaluation data collected.

Another way to analyze evaluation data is to hire an external evaluator with expertise in data analysis and access to statistical software. While costs may be involved for both the purchase of software and the hiring of external evaluators, your analysis can be done more efficiently, and an external evaluator can provide expertise to ensure the data analysis is executed appropriately and accurately reflects program outcomes. See the link on page 26 of this toolkit for a list of AEA members available for consulting on evaluations.

Discussion Points for Analysis
The individuals analyzing the evaluation data—the “analysts”—need to address several questions to frame the analysis. It is very helpful for those sponsoring the evaluation, such as the SEA and LEA administrators, to be involved in addressing these questions, as the decisions will shape the usefulness of the findings. A number of these questions, such as identifying the most pertinent outcomes and the level at which outcomes should be measured, were discussed earlier. Here, we add two issues that have to do with the data analysis.

- What does missing information mean? What information, and on whom, is missing? How will missing information affect analysis?
- What are common issues in the analysis of education-related data that we need to consider (e.g., clustering of students in schools, student mobility)?

What does missing information mean? What information is missing?
The fact that some data are missing may or may not be meaningful. You may ask the analysts to evaluate the effectiveness of numerous principal preparation programs on placing candidates into high-need schools within two years of completing the program. In this example, it is possible that some participants in the principal preparation program will not have information about the schools in which they were placed at the two-year mark. On one hand, there is meaningful missing information: Not having data on where the participant is placed indicates she did not get a job as a principal within two years of program completion. On the other hand, there is missing information that may not relate to the analysis. Perhaps you did not have sufficient resources to track participants after they left the preparation program, so it is unknown whether the person got a job in a high-need school, got a principal job at a different type of school, or did not receive a principal position within two years. Missing data has the ability to affect the analysis and outcomes that can be identified from an evaluation. Ensuring that the analysts understand the information they receive in the various databases can help ensure the available information is being utilized for the impact evaluation.

As the analysts begin to work with the data, one of the first practices they should take on is exploring what information is missing and how much information is missing in the data. They might also look at whether the data are missing from both the intervention and comparison groups—if one group has more or different missing information, the difference might reflect a real and meaningful difference between the groups’ experiences. A conversation with the analysts may help identify ways fellow evaluators can remedy problems with the most critical missing pieces of data. These conversations should happen with each additional database or year of data collected for the analysis.

Some statistical software packages, such as R (https://www.r-project.org/about.html), are available free and have substantial public documentation and support available. Users need to have, at minimum, basic data handling and statistical programming knowledge to use the program.
What are common issues in the analysis of education-related data that we need to consider (e.g., clustering of students in schools, student mobility)?

While research designs (e.g., QEDs) used for education evaluations are the same methods that can be used to study other topics—housing markets, new drug therapies, or even professional athletes—there are some characteristics of education that call for special analysis strategies.

Students within a school are much more similar to each other than to students in other schools. Prior research has shown that families and students of similar backgrounds and socioeconomic status tend to enroll in schools together. Once those students are in a school, common instructional and environmental factors affect them. Students directly affect each other as they interact in classrooms, the hallway, and on the playground. These factors make their educational outcomes more similar in ways unrelated to the intervention being studied. The same phenomenon occurs with teachers. Teachers in a school often have common backgrounds and credentials. Once in a school, teachers are exposed to the same administrators, professional development programs, and other environmental factors. Given the connectedness of individuals in a school, interventions are likely to affect them in a similar way. Analysts familiar with these data (and human) relationships will typically consider the “clustering” of data. Clustering refers to the natural grouping of individuals or observations, such as a school or classroom, which may cause their outcomes to be correlated, or connected. Clustering can affect the analysis if the clusters aren’t accounted for statistically. The WWC describes the issue and provides a solution in the “Clustering Correction for ‘Mismatched Analyses’” section of the WWC Standards and Procedures Handbook (U.S. Department of Education Institute of Education Sciences, 2017a). By accounting for the data clusters, the evaluation results are likely to better reflect the true effect of the evaluation.

Mobility, or student movement into and out of schools, is another education-specific concern. Students leave a school when they graduate to the next school level or move to another area, for example. They may transfer from a school participating in the intervention to a school not participating, or vice versa. Teachers may also move from or to a school during the study. If the intervention group contains students or teachers who were not exposed to the intervention, or the comparison group contains students or teachers who were, the outcomes of the analysis may not truly reflect the impact of the intervention. Further, losing students or teachers from the study (because they are no longer in a studied school) can make the analysis less powerful and possibly biased. Analysts need to account for the movement of people across places (see the box in Chapter 3 titled “Tip: Well-Designed and Well-Implemented Studies” for more information on attrition).

Analysts familiar with education data may already know to look for education-specific challenges. Analysts new to education data may need more time to work with the data, exploring these issues and asking questions about state or local context to ensure the statistical analyses are conducted accurately. SEA and LEA administrators can also help analysts by providing some guidance on issues that are unique to their context or that have been challenges in other evaluations. The more information analysts have, the better the results of an evaluation.

Summary
You should now have a clear understanding of
- how to organize evaluation data
- ways to analyze the evaluation data
- points that should be discussed with the analysis team.

To best utilize the scarce resources available for evaluation purposes, it helps to be clear about your data needs before rolling out your intervention and revisit those data needs as the program evolves. Thinking
through the information needed for evaluation early also ensures that the lessons learned are useful and actionable.

In the next chapter, we provide more guidance on how to put evaluation information to good use to inform future program operations and success.
Chapter 6: Using Your Evaluation Data to Improve Your Program

In this chapter, we first guide you in assessing your evaluation to better understand factors that may affect the accuracy of your findings. We also review how to interpret evaluation findings to determine the extent to which the program achieved its intended outcomes and conclude with a few tools your evaluators can use to identify appropriate changes to improve the quality of the program. These steps should be taken once the evaluation data have been analyzed and results are available.

Assess Your Evaluation

Before interpreting and applying the evaluation data, revisit the final evaluation design as it was actually implemented to determine

- how well the actual program participants reflect the intended program participants
- how well the evaluation participants reflect the program participants
- how well the program implementation aligned with the program logic model.

Information on who participated in the program, who was included in the evaluation, and how well those two groups aligned helps inform decisions about the quality of the evaluation conducted and how to interpret results of the statistical analysis. If a process or implementation evaluation was also conducted, these results can also bolster decisions about the types of program improvements needed.

Let us say, for example, that 90 percent of intended participants actively attended or participated in the intervention, and 90 percent of those active participants were included in the outcomes evaluation. Because the proportion of participants reached and included in your evaluation was relatively high, your evaluation represented the intended population well. This helps provide confidence in the results of the outcomes identified.8 Evaluation results may not accurately describe the effectiveness of your program if they do not or only somewhat represent the intended population. Future evaluations will require intentional program recruitment and retention strategies to ensure that the necessary number of participants is engaged in each phase of the intervention.

The evaluators may want to consider how active participants differ from participants who were not active in either the intervention or the evaluation. If key demographic groups of intended participants are not in the active groups, it is possible results may not accurately reflect how the intervention would serve all school leaders. In evaluations using experimental designs and QED, these differences would be accounted for and adjusted, where appropriate, in the analysis phase to ensure your results accurately reflect the program’s impact.

Finally, revisiting the logic model helps verify that the intervention or program was implemented as intended. The assumed resources were those utilized, the activities were delivered as intended (e.g., highly qualified principal candidates were recruited into the program), and the intended outputs from the program were produced. Deviations from expectation may help explain potential barriers to meeting the intended outcomes. Simultaneously, these deviations may also explain successes of the program in meeting intended targets. Documenting these differences is critical for decisionmaking, as well as for sharing your evaluation results with stakeholders. Moreover, districts and states looking to adopt new school leadership interventions may look to your results to inform their decisions. Being explicit about deviations will help them identify ways the program may or may not work in their contexts.

---

8 See the WWC Procedures and Standards Handbook, V4.0, (U.S. Department of Education Institute of Education Sciences, 2017a, p. 11) for guidance on acceptable levels of attrition.
Interpret the Evaluation Data
With confidence in the results of your evaluation, the next step is to interpret evaluation findings by assessing whether they provide useful information about the extent to which the program achieved its intended outcomes.

Tool 6.1 provides an example of how the evaluator can document performance on the pertinent outcomes, whether the desired outcomes were achieved, and whether those changes fell short of expectation, met expectations, or exceeded expectations. Evaluators should also decide, based on those results, whether action is needed to improve the program or address additional needs of school leaders. As discussed in Chapter 4, creating expectations or benchmarks for intervention success prior to receiving evaluation results will help with interpreting the findings documented in this tool. The final column of Tool 6.1 asks the evaluator to consider any contextual factors or implementation challenges that may have contributed to any expectations that were not met. This information may help shape any action taken to improve the program for future participants. Importantly, recommendations for program improvement or changes to practice or policy should be informed by the collective body of information documented in Tool 6.1 and not the results from only one outcome measure.
### Tool 6.1: Review Program Changes, with an Example

<table>
<thead>
<tr>
<th>Outcome of Interest</th>
<th>Desired or Expected Change</th>
<th>Change in Outcome Identified</th>
<th>Outcome Direction</th>
<th>Met Expectations?</th>
<th>Connection to Program Activities/Outputs</th>
<th>Action Needed?</th>
<th>Potential Barriers (e.g., resources, expertise)?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Short-term, intermediate, or long-term outcome measured</td>
<td>Factor and amount by which the program was expected to affect the outcome</td>
<td>Evaluation outcome</td>
<td>What is the direction of the outcome (e.g., better, worse)?</td>
<td>Did this meet your expectations for the program?</td>
<td>How were program activities or outputs related to program performance on this outcome?</td>
<td>Does the SEA or LEA need to take more action to meet this desired outcome?</td>
<td>What may have prohibited the intervention from creating desired changes? What contextual factors or implementation challenges shaped this result?</td>
</tr>
<tr>
<td>Outcome measure</td>
<td>Factor: Amount: Instrument:</td>
<td>Check one</td>
<td>Better ☐ Same ☐ Worse ☐</td>
<td>Check one</td>
<td>Fell Short ☐ Met ☐ Exceeded ☐</td>
<td>Text:</td>
<td>Check one</td>
</tr>
<tr>
<td>Example: School climate</td>
<td>Factor: average school climate rating</td>
<td>Amount: 15%</td>
<td>□ Better</td>
<td>□ Fell Short</td>
<td>Only 70% of placed participants received ongoing training from their district on staff hiring, selecting professional development opportunities for teachers, and providing support for addressing instructional quality.</td>
<td>□ Yes</td>
<td>Districts have different levels of financial capacity to support principals after placement.</td>
</tr>
</tbody>
</table>

Example: School climate  
Factor: average school climate rating  
Amount: 15%  
Instrument: teacher surveys  
Amount: 20% increase  
Instrument: teacher surveys  
Amount: 15%
**Share the Evaluation Results**
Sharing the results of the evaluation with SEA and LEA leadership, stakeholders, community members, and program participants can and should occur as part of any strong evaluation. Determining the specific timing of this reporting may depend on the expectations of program and SEA and LEA leaders, the results of the evaluation, or the overlap with critical program or policy time frames (e.g., applying for ESSA program funds that require demonstrated evidence). In some cases, SEA and LEA administrators may want to delay the sharing of results until a plan for how unmet outcomes will be addressed in future program cycles is prepared to accompany concerning results. By contrast, SEA and LEA administrators may see the need for rapid communication of evaluation findings to ensure consistent or additional funding can be secured for future program operations. Regardless of which approach SEA and LEA administrators opt to take, dissemination of results is important for numerous reasons: The reporting may create publicity that reaches the program’s target population and assists in future program recruitment; sharing with key stakeholders may help garner the additional resources and supports identified as reasons the program did or did not meet expected outcomes; and the results may help shape policy and practice that inform future operation of the programmatic efforts (Centers for Disease Control and Prevention, 2009).

**Summary**
You should now have a clear understanding of
- how the quality of your evaluation may impact your interpretation of your findings
- how to interpret the findings from your evaluation, including whether goals were met and what future actions may need to be taken
- the importance of sharing evaluation results with key policymakers and administrators.
Appendix A: Existing School Leadership Interventions with Documented Evidence

Tool A.1 summarizes the evaluation design and outcome measures used in the studies reviewed in the School Leadership Interventions report that met Tiers I–III evidence requirements and provided positive findings. These studies focused on six types of school leadership programs: principal preparation programs, professional learning, leader evaluation systems, strategic staff management, working conditions, and leadership-focused school reform models. The majority of the research is based on QEDs, and only two program types (i.e., professional learning and school reform models) used experimental designs with random assignment of participants (usually aspiring principals or students) into an intervention or nonintervention condition. Appendix D in the School Leadership Interventions report provides a one-page, detailed description of each of the interventions referenced in Tool A.1.

As seen in the toolkit, experimental designs and QEDs can be analyzed using numerous statistical techniques. A common element of the QED evaluation approaches listed in the toolkit is the use of rigorous methods to identify a comparison group: matching on characteristics prior to the intervention. Once that comparison group is identified, evaluators use the statistical approach that best fits the available data and that can identify changes or differences in the outcomes of interest.

The most frequently utilized outcome measure in these prior studies of school leadership interventions is student achievement as measured by annual test scores. Some of the evaluation or research teams have also examined program impacts on teacher and principal retention or turnover, principal leadership behaviors, as well as student behaviors and attitudes. Most outcome measures are available from administrative data already collected each year by states and districts. Surveys administered to students, staff, or parents typically require additional resource expenditure by either the evaluators or the SEA or LEA utilizing the intervention program (e.g., state departments of education). The outcome measures are discussed in more detail in Chapter 4 of this guide.
### Tool A.1. Studies of Leadership Interventions

<table>
<thead>
<tr>
<th>Program Type</th>
<th>Intervention Activities</th>
<th>Study</th>
<th>ESSA Tier</th>
<th>Analysis Method* (study design)</th>
<th>Outcomes (as measured by)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Principal preparation</td>
<td>Principal preparation programs: Usually include graduate-level courses, internship working with or as principal, and summative test</td>
<td>Fuller, Young, and Baker, 2011</td>
<td>Tier III</td>
<td>Ordinary Least Squares regression (correlation analysis with controls)</td>
<td>Student achievement (test scores)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Quality of teaching staff (alignment between certification status and subjects taught, from personnel records)</td>
</tr>
<tr>
<td>New Leaders: Preparation program involves selective recruitment and admissions, training, and endorsement and support for principals early in their tenures</td>
<td>Gates et al., 2014</td>
<td>Tier II</td>
<td>Student fixed effects (QED) Nearest neighbor matching (QED)</td>
<td>Student achievement (test scores)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Principal working conditions and time use (self-report surveys)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Principal retention (years remaining in same school, from personnel records)</td>
</tr>
<tr>
<td>Texas Principal Excellence Program: Involves seminars and workshops on leadership competencies and related topics, accompanied by feedback and guidance</td>
<td>Fouche, 2011</td>
<td>Tier II</td>
<td>Pre-/postintervention differences (correlation analysis without controls, Tier IV evidence) Propensity score matching (QED)</td>
<td>Leadership behaviors (participant self-reports and ratings of peers on 360-degree leadership evaluation survey)</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>School performance (proficiency rate)</td>
</tr>
<tr>
<td>Professional learning</td>
<td>McREL Balanced Leadership Program: professional development focused on purposeful communities, research-based practices, and assessing needs</td>
<td>Jacob et al., 2015</td>
<td>Tier I</td>
<td>Random assignment of schools to intervention and control (RCT), pre- and postintervention differences</td>
<td>Principal Leadership Practices (McREL-aligned survey of leadership qualities)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Teacher and principal turnover (personnel records)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Student achievement (test scores)</td>
</tr>
</tbody>
</table>

*We do not define these statistical analysis procedures here but recommend that state and local leaders consider the statistical approach that best fits their intervention, sample, and outcomes, using guidance from a statistical expert if helpful.*
<table>
<thead>
<tr>
<th>Category</th>
<th>Description</th>
<th>Study</th>
<th>Tier</th>
<th>Methodology</th>
<th>Outcome measures</th>
</tr>
</thead>
<tbody>
<tr>
<td>NISL Executive Development Program</td>
<td>professional development focused on vision and goals, teaching and learning, and sustaining improvement</td>
<td>Nunnery et al., 2011</td>
<td>Tier II</td>
<td>Repeated measures analysis of variance (ANOVA) with matched comparison group (QED)</td>
<td>School performance (test scores aggregated to single school-average measure)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Nunnery, Ross, and Yen, 2010</td>
<td>Tier II</td>
<td>Factorial ANOVA with matched comparison group (QED)</td>
<td>Student achievement (test scores)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>School performance (proficiency rates)</td>
</tr>
<tr>
<td>Working conditions</td>
<td>Conditions designed to improve principal effectiveness, such as incentives and autonomy</td>
<td>Abdulkadiroglu et al., 2011</td>
<td>Tier II</td>
<td>Instrumental variable analysis (QED)</td>
<td>Student achievement (test scores)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Steinberg, 2014</td>
<td>Tier II</td>
<td>Regression discontinuity design (QED)</td>
<td>Student achievement (test scores)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>School performance (proficiency rates)</td>
</tr>
<tr>
<td>Leadership-focused school reform models</td>
<td>KIPP: a public charter school network. “Power to lead” is one of the five pillars, or operating principles. Core elements are development, autonomy, and accountability</td>
<td>Angrist et al., 2012</td>
<td>Tier I</td>
<td>Random assignment of students to intervention and control (RCT), pre- and postintervention differences Instrumental variable analysis (QED)</td>
<td>Student achievement (test scores)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gleason et al., 2014</td>
<td>Tier II</td>
<td>Propensity score matching (QED)</td>
<td>Student achievement (test scores)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Student achievement gains (test score changes in subsequent years)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Gallagher and Ross, 2005</td>
<td>Tier II</td>
<td>ANOVA with matched comparison group (QED)</td>
<td>Student achievement (test scores)</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Tuttle et al., 2010</td>
<td>Tier II</td>
<td>Propensity score matching (QED)</td>
<td>Student achievement (test scores)</td>
</tr>
</tbody>
</table>

10 The EDGAR guidance and WWC Handbook were modified since the original study, to explicitly consider RDD studies as “experimental” and therefore eligible for Tier I. At the time of the study, RDD studies were considered eligible for Tier II evidence.
<table>
<thead>
<tr>
<th>Study</th>
<th>Tier</th>
<th>Intervention Method</th>
<th>Outcomes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Woodworth et al., 2008</td>
<td>Tier II</td>
<td>Matched sample</td>
<td>Student achievement (test scores)</td>
</tr>
<tr>
<td>Tuttle et al., 2013</td>
<td>Tiers I, II</td>
<td>Random assignment of students to intervention and control (RCT), pre- and postintervention differences</td>
<td>Student achievement (test scores)</td>
</tr>
<tr>
<td>Tuttle et al., 2015</td>
<td>Tiers I, II</td>
<td>Random assignment of students to intervention and control (RCT), pre- and postintervention differences</td>
<td>Student achievement (test scores)</td>
</tr>
<tr>
<td>UVA School Turnaround Specialist Program:</td>
<td>Tier II</td>
<td>Difference-in-differences (QED)</td>
<td>Student achievement (test scores)</td>
</tr>
<tr>
<td>Involves a planning year with the district, school-leadership-selection support, and executive development for school leaders and turnaround teams in residential programs and on-site coaching</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Player and Katz, 2016</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Appendix B: Implementation Evaluation and Process Measures

While the focus of this toolkit is to help guide your selection of an outcome evaluation method, particularly for SEAs and LEAs using Tier IV interventions, it is also important to consider implementation or process inputs—which focus on the resources, activities, and outputs from the logic model. Effective implementation of a program is crucial for that program to meet its intended outcomes. Conducting an evaluation that incorporates both implementation and impact components allows evaluators, administrators, and policymakers to (1) understand the factors that may or may not influence a program’s ability to produce intended outcomes, (2) identify ways to continuously improve ongoing programs, and (3) act as an informative resource for other adopters of the same or similar interventions.

In some cases, such as highly structured or scripted interventions, effective implementation involves strict adherence to the program as designed; impact of the intervention may be lower if the intervention is not implemented as planned. In other cases, effective implementation involves adaptation of the program to local conditions; intentional variations in the intervention may be necessary for the context and may increase impact. Rigorous documentation of the implementation process, including the extent to which an intervention was adhered to or modified, can inform decisions of both current and future implementers. Results of implementation evaluations can also help you shape decisions such as the allocation of resources or where program activities may need to be supplemented to better reach the target population or produce intended outcomes.

Implementation is a multifaceted process and can be explained by the following eight dimensions (Humphrey et al., 2016).

1. Fidelity, or the extent to which implementers adhere to the intended program model.
2. Dosage, or how much of the intended intervention has been delivered and/or received.
3. Quality, or how well different components of an intervention are delivered to participants.
4. Reach, or the rate and scope of participation among intended and unintended groups.
5. Responsiveness, or the degree to which participants engage with the intervention.
6. Program differentiation, or the extent to which intervention activities can be distinguished from other, existing practice.
7. Monitoring of control/comparison groups, or otherwise measuring what happens where no intervention is being implemented.
8. Adaptation, or the extent to which changes to the intervention were deliberately made during delivery.

Implementation or process evaluations seek to answer such questions as

1. Was the program implemented with fidelity to the intended model?
2. How much or what dosage of the intervention did participants receive?
3. Did participants actively engage in the activities being delivered? Did they demonstrate understanding of the program and its purposes?
4. What contextual factors shaped implementation or program delivery in each participating site?
5. To the extent that implementation deviated from the original intent, to what extent was that intentional adaptation? To what extent was it unintended?
6. What other resources would have improved the delivery of the intervention?

The indicators or measures used for implementation evaluations also differ from those of outcomes and impact evaluations. Example measures of implementation include

1. proportion of program implemented as designed (e.g., three of four components implemented as designed; all four components partially implemented)
2. characteristics of staff delivering the intervention and participants served (e.g., all trainers were experienced principals; participants were first-year principals)
3. the number of participants served and the dosage received (e.g., total hours, number of sessions attended, or proportion of whole intervention participant engaged in)
4. number of intervention activities provided, and when they occurred (e.g., training occurred monthly at the district office, for four hours on Saturdays)
5. proportion of participants accurately describing intervention activities and purposes after participation
6. resources used and resources needed: financial, time, technological (e.g., 15 laptops).

For an example of implementation data collected on a school leadership intervention, see Jacob et al., 2015, pp. 325–326. These evaluators collected information on the training and consistency of data collectors, expertise of the facilitators delivering the intervention to principals, the average amount of the intervention received by participants, and the participant’s satisfaction with the professional development offered in each program session.

Finally, it is also important to consider whether drivers of strong implementation are in place: communication across stakeholders, processes for identifying and addressing challenges, supports for implementers, leadership engagement, and data systems.

Make a Plan to Improve Your Program
As mentioned previously, two benefits of conducting a complementary implementation evaluation to an outcome evaluation are that implementation measures can help identify why a program did or did not meet the intended outcomes, as well as help identify ways to improve the program’s future operation and outcomes.

Once the evaluation data have been analyzed and the evaluators identify where the program did or did not meet expected outcomes, the evaluators, along with key stakeholders from the SEA or LEA, can begin to identify what efforts are needed to improve the intervention outcomes in the future. The review of the evaluation results could suggest different actions. Tool B.1 presents a series of results-based scenarios and associated strategies for program improvement. The scenarios are based on whether your program was implemented effectively and whether the program achieved the expected outcomes.

Data from an implementation evaluation can tell you whether program activities were or were not implemented with adequate dosage (i.e., participants got most or all of the intervention) and fidelity (i.e., program content was delivered as intended). The implementation evaluation can also identify whether the target population participated in the intervention or whether most of the intervention was delivered to non–target population individuals.

Tool B.1 guides you through a quick interpretation of the evaluation data. Based on what is concluded after consulting Tool B.1, use that information in implementing Tool B.2 (adapted from Acosta et al., 2013). For example, if your review indicated that the program did not meet expectations (i.e., “fell short”) in Tool B.1, the program activities did not result in a significant change, or at least the desired change, on intended outcomes. Identify which scenario in Tool B.2 best describes your evaluation results. Then, proceed to the small-scale Continuous Quality Improvement (CQI) assessments in Tool B.3, which will help you to select improvement strategies, if needed.
### Tool B.1. Results-Based Scenarios and Associated Strategies for Program Improvement (Acosta et al., 2013)

<table>
<thead>
<tr>
<th>Program activities resulted in a significant change on intended outcomes</th>
<th>Yes</th>
<th>No</th>
</tr>
</thead>
</table>
| Yes                                                                   | The program seems to be working as designed; continue implementation and evaluation.  
  - Keep monitoring process and outcomes.                                | Changes may be needed because your program did not achieve the intended results. Lack of significant results could be related to the following barriers:  
  - a small sample size  
  - a misalignment between program activities and intended outcomes  
  - a mismatch between the program and its participants (e.g., participants started high on outcome measures, leaving limited room to improve)  
  - a mismatch between evaluation design or measures and program outcomes.  
  Address these potential barriers before implementing the program. Addressing these barriers may require you to change your program activities and your evaluation design/measures. |
| No                                                                    | Changes may be needed to address potential barriers to fidelity and dosage; you might want to reevaluate the program to determine whether results improve.  
  - Focus on strengthening areas that were not implemented with fidelity.  
  - Improve recruiting or retention strategies to ensure that participants get an adequate dosage.  
  - Consider whether adaptation to context is necessary for this program. | Changes may be needed because the program does not seem to be working as designed. Significant changes to program activities and evaluation design may be required. Assess whether  
  - there is a mismatch between the program and the intended population  
  - there are adequate resources to deliver the program (e.g., Do facilitators have enough training? Do participants have incentive to attend program?).  
  Proceed to Tool 6.1. |

If your review suggests that changes might be needed, you can use the following CQI assessment to identify the specific changes needed and to plan key activities (see Tool B.2). If you answer “no” to any of the questions about potential CQI challenges, review the suggested improvement actions to address challenges and identify those feasible for your program to pursue.
Tool B.2. What CQI Actions Are Needed to Improve the Program? (Acosta et al., 2013)

<table>
<thead>
<tr>
<th>Questions About Potential CQI Challenges</th>
<th>If No, Consider These Improvement Actions to Address Challenges</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Did participants represent your target population?</td>
<td>Review eligibility and outreach criteria to ensure that they are clear enough to recruit appropriate program participants. Review screening criteria and processes.</td>
</tr>
<tr>
<td>□ Yes</td>
<td></td>
</tr>
<tr>
<td>□ No ➤</td>
<td></td>
</tr>
<tr>
<td>B. Was the program delivered as intended?</td>
<td>Improve program management and staff training on how to implement the program and self-assess fidelity.</td>
</tr>
<tr>
<td>□ Yes</td>
<td></td>
</tr>
<tr>
<td>□ No ➤</td>
<td></td>
</tr>
<tr>
<td>C. Was attendance adequate?</td>
<td>Assess whether there are any logistical barriers that might make it difficult for participants to participate. Consider whether changing the time or place of the program would improve participation. Consider whether other program delivery options might be more appropriate for the population served.</td>
</tr>
<tr>
<td>□ Yes</td>
<td></td>
</tr>
<tr>
<td>□ No ➤</td>
<td></td>
</tr>
<tr>
<td>D. Did you have the resources needed to implement the program completely and as intended?</td>
<td>Review your program’s resources for implementation and evaluation to determine whether you have the right staff, resources, and partnerships to deliver the program. Consider other resources.</td>
</tr>
<tr>
<td>□ Yes</td>
<td></td>
</tr>
<tr>
<td>□ No ➤</td>
<td></td>
</tr>
<tr>
<td>E. Were the outcomes you expected reasonable/appropriate for the program?</td>
<td>Revisit the goals and logic model that you developed and revise them to be more reasonable/appropriate for your program. Analyze data to explore possible reasons why outcomes failed to meet expectations.</td>
</tr>
<tr>
<td>□ Yes</td>
<td></td>
</tr>
<tr>
<td>□ No ➤</td>
<td></td>
</tr>
<tr>
<td>F. Was your process and outcome evaluation appropriate?</td>
<td>Update the process and/or outcome evaluation plan to be more appropriate for your program while maintaining the most rigorous and valid research design feasible.</td>
</tr>
<tr>
<td>□ Yes</td>
<td></td>
</tr>
<tr>
<td>□ No ➤</td>
<td></td>
</tr>
</tbody>
</table>

Transfer the relevant improvement actions from any items on Tool B.2 for which you answered “No” to Tool B.3. Then, record who will participate in the action, who will be responsible for the action, the resources needed, location details, and the target date for improvement. Making a plan for program improvement using Tool B.3 will help you identify the actions necessary to achieve those objectives and specify a target date for completion of program improvement activities. If possible, complete the program improvement activities prior to implementing the program again.
<table>
<thead>
<tr>
<th>Improvement Action</th>
<th>Who Will Participate</th>
<th>Who Is Responsible</th>
<th>Resources Needed/Source</th>
<th>Location/Details</th>
<th>Date of Completion</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example: principal incentive program</td>
<td>District superintendent</td>
<td>District superintendent</td>
<td>Additional funding</td>
<td>Discussion and plans</td>
<td>August 20xx, before beginning of next school year</td>
</tr>
<tr>
<td>Increase principal incentives to a motivating level as acknowledged by district and school leaders</td>
<td>Principal supervisors</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Board of Education members</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
More information about implementation evaluations:


References


Nunnery, J. A., S. M. Ross, S. Chappell, S. Pribesh, and E. Hoag-Carhart, The Impact of the NISL Executive Development Program on School Performance in Massachusetts: Cohort 2 Result, Norfolk, Va.: Center for Educational Partnerships at Old Dominion University, 2011.


https://www.issuelab.org/resources/11029/11029.pdf?download=true

WWC—See U.S. Department of Education Institute of Education Sciences, What Works Clearinghouse, 2017