Program Evaluation Toolkit Module 3, Chapter 1: Evaluation Design

Regional Educational Laboratory Central

From the National Center for Education Evaluation at IES

Speaker 1:

Welcome to the first chapter of module 3 of the program evaluation toolkit. In this module, you will learn major considerations for designing an evaluation.

With an understanding of how a logic model frames your evaluation and a sense of what types of questions you wish to answer, you can consider how the design of your evaluation is crucial to achieving its goals. This module will help you select a design aligned to your evaluation questions. The questions you wish to answer largely influence the evaluation design you choose. However, real-world constraints such as timeliness and evaluation team capacity, as well as practical considerations such as the ability of evaluation partners to implement certain procedures, also inform your selection of an evaluation design.

After completing this module, you should have a basic understanding of (1) different evaluation design options, (2) some of the ways to ensure that claims made from an evaluation are justifiable, and (3) the different tiers of evidence provided by different evaluation designs. The AMMP! (After-School Middle-Grades Math Program) example from module 1 will help illustrate these principles. It would be helpful to have the AMMP! logic model and research questions at hand as you progress through module 3.

This module includes three chapters, each highlighting different aspects of evaluation design.

Chapter 1 provides an overview of different evaluation design categories and considerations.

Chapter 2 reviews threats to validity that you should consider when designing an evaluation.

Chapter 3 presents the Every Student Succeeds Act tiers of evidence and the What Works Clearinghouse evidence standards, which help inform the development of evaluation designs.

Refer to the resources page of the website for worksheets, templates, and other resources to help you work through this module.

Let’s get started with the first chapter, “Evaluation Design Options and Considerations.”

Evaluation design refers to the data collection processes and analytic methods used to answer your evaluation questions.

An evaluation design should be informed by the program goals, logic model, evaluation questions, available resources, and funding requirements.

In this chapter, you will learn about four broad categories of evaluation design: descriptive designs, correlational designs, quasi-experimental designs, and randomized controlled trials.

Often, program evaluation involves more than one, or even all, of these design categories. Programs often begin as pilot programs, which may be small in scope. If promising evidence is found using descriptive or correlational designs, those programs may be scaled up. To assess whether programs are effective after the initial pilot, more rigorous designs—such as quasi-experimental or experimental designs—may be used.

Let's begin by examining what each of these evaluation design categories entails. As each design category is introduced, you will explore whether the design might be used in the case of AMMP!, and you will be prompted to consider what the design might look like if AMMP! were scaled up. Next, you will learn how an evaluation may involve multiple design categories.

First let's discuss descriptive designs. As the name implies, this design category involves describing a program by addressing "who," "what," "where," "when," and "to what extent" questions as they relate to the program. A descriptive design can be used to document how a program works, provide feedback on implementation, identify barriers to program success, help determine the best outcomes for assessing program effectiveness, or help clarify program objectives.

To get a better sense of what a descriptive evaluation might look like, let's review the AMMP! example.

For the AMMP! evaluation, let's say that the team wants to document the number of volunteer tutors trained to ensure that there are enough staff to provide after-school homework support to students. The team might even want to know the backgrounds of the students and volunteers who participate in the program. The questions presented here might provide insight into key activities and outputs.

Example descriptive design questions include the following:

- How many volunteer tutors were trained to implement AMMP!?
- How many tutoring hours, on average, did students receive?
- What are the characteristics of students and volunteers participating in the program?

A descriptive design answers descriptive questions. The descriptive design for AMMP! will help the evaluation team understand whether enough tutors were trained, whether those tutors are meeting with students, and whether the target population of students is being reached.

Next, let's look at correlational designs. A correlational design involves identifying a relationship between two variables and determining whether that relationship is statistically meaningful. Correlational analyses do not demonstrate causality. You might learn that X is related to Y, but you cannot confirm that X caused Y in a correlational design. A correlational design can be used to document how program participation relates to outcomes of interest; understand associations between various subgroups and changes in participants' knowledge,

skills, and behaviors; or determine how differences in implementation are associated with intended outputs.

What about an example of a correlational design? For the AMMP! evaluation, the team likely wants to know whether there is a relationship between tutors' participation in the AMMP! training, one of the logic model activities, and their knowledge of effective techniques, one of the logic model short-term outcomes. They might ask the following correlational design question:

- Did the tutors who participated in AMMP! training demonstrate an increased knowledge of effective techniques?

However, the team will not be able to say that the training caused increased tutor knowledge. A common mistake is using descriptive or correlational designs to say, "this intervention caused this to change."

Now let's turn to two evaluation designs that can be used to determine if a program is likely to have caused its expected outcomes: quasi-experimental designs (or QEDs) and randomized controlled trials (or RCTs). Both QEDs and RCTs compare the outcomes of a treatment group, who receive the intervention, to the outcomes of a comparison group, who do not receive the intervention. Consequently, both QEDs and RCTs are often used to answer the same types of evaluation questions. Including a comparison group helps ensure that any changes observed in the treatment group are actually due to the intervention and not to some other cause.

In QEDs, individuals are not randomly assigned to groups because of ethical, practical, or other constraints. For example, the AMMP! funder may require that all students interested in the program be allowed to participate. Because there may be systematic differences between students who participate in the program and those who don't, QEDs use matching or other statistical adjustments to create treatment and comparison groups that are as similar as possible to one another.

RCTs involve randomization, a process like a coin toss, to assign individuals to the treatment or comparison group. Using this process virtually ensures that there are no systematic differences between individuals in the treatment and comparison groups. Because of this, RCTs are often referred to as "the gold standard" for evaluation and are typically held to offer the highest-quality evidence of effectiveness of a program.

Both QEDs and RCTs can be used to compare differences in outcomes between treatment and comparison groups. They can be used to determine whether a program or intervention is likely to have increased intended outcomes of participants on average, or to evaluate how confident an evaluation team can be that a program or intervention has caused changes in outcomes that matter to the team. Let's look at an example of each to help illustrate the differences.

First, an example of a QED. What if the AMMP! evaluation team wants to understand whether the program has had an effect on students' readiness for high school math? In this case, the team can stretch this example to include a set of partner schools that serve similar populations of students. Because these schools are not using AMMP!, they provide a potential comparison

group. These schools are using a "business as usual" condition in which no new interventions or programs are offered.

However, the partner schools that are not using AMMP! may differ from the AMMP! schools in many ways. For instance, they may differ in the characteristics of the students they serve (such as the percentage of students with individualized education programs or the average student achievement), the characteristics of teachers (such as the average years of teaching experience), or other factors (such as average teacher salary). To begin the process of designing a high-quality evaluation, the AMMP! evaluation team should gather data about the student and teacher populations at the schools. The team can use these data to begin to understand how similar the AMMP! and non-AMMP! schools, students, and teachers are to one another, across a number of factors. The team can also use the data in statistical models to adjust, at least in part, any differences between the treatment and comparison groups that might be related to the outcomes the team is most interested in. This design will help the team address the mid-term outcome about readiness for high school math.

An example QED question is:

- Did participation in AMMP! have an effect on grade 9 students' readiness for high school math as measured by a math placement test?

A QED like this one is appealing because the evaluation team can use existing data from schools that did and did not implement AMMP! to answer the evaluation question. In comparison, an RCT would require identifying a set of schools that agree to be randomly assigned to either implement or not implement AMMP!. However, as you will explore further in chapter 3, evidence from a well-designed, well-executed QED is not as strong as evidence from a successful RCT.

Now for an RCT example. Let's suppose that the AMMP! evaluation team has the funding and support to roll out the program to only a limited number of students in a subset of schools. Students in some schools will be randomly assigned to use AMMP!, and students in other schools will be randomly assigned to continue with the "business as usual" condition in which no after-school program is offered. This design, if well-executed, generates the strongest possible evidence for the team as it works to evaluate whether AMMP! participation has decreased the number of issues in the community.

An example RCT question is:

- Did participation in AMMP! cause a decrease the number of community issues among students in the program?

An RCT like this one can help the evaluation team understand whether AMMP! participation, in addition to other factors, caused a decrease in the number of community issues. In this way, the team can feel very confident that the evaluation results are due to AMMP! and not to other unobserved factors that might influence the decrease in community issues.

Importantly, because of random assignment, the students who participate in AMMP! and the students who don't participate in AMMP! are likely to be similar in observable ways, such as student achievement, as well as in unobservable ways, such as family support. In a QED, unobservable differences are more likely. For instance, students who participate in AMMP! might have more family support than students who don't participate in AMMP!. Such support is difficult to measure but influences student outcomes. For this reason, a well-designed RCT gives the evaluation team stronger evidence of the causal effects of AMMP! than a QED does.

One word of caution: Participants who are assigned to a treatment group may not always actually participate in an intervention as some interventions are optional. For instance, AMMP! is a program at the school and is optional for students to participate. All eligible students at the school are still assigned to the treatment group, but some may choose to not participate in AMMP!. It is important to account for this in an analysis. The simplest way to do so is to look at the effect of "intent-to-treat," where all students assigned to the intervention are examined, regardless of whether they participated in the intervention. Another way to look at the effect of "treatment on the treated," where all students who actually did or did not participate in the intervention are examined.

To test your knowledge of the different design categories, complete *Evaluation Design: Matching Activity*, available on the resource page of the website. Pause the video now to complete this activity before we discuss the correct matches.

OK, let's review the activity.

The answer to example 1 is C, or a quasi-experimental design. The key here is that the treatment group and the control group are statistically similar, but random assignment was not used to determine the groups.

The answer to example 2 is A, or a descriptive design. There are no treatment and comparison groups, and the analyses are limited to assessing frequencies and perceptions related to the online program.

The answer to example 3 is B, or a correlational design. Here the administrators examine how one variable—student-to-counselor ratios—relates to other variables—attendance and college-bound students. This approach takes the evaluation a step further than a descriptive design, but again there are no treatment and comparison groups.

Finally, the answer to example 4 is D, or a randomized controlled trial. In this case, the district conducts a lottery to determine which schools receive the intervention. Thus, the evaluation includes randomly assigned treatment and comparison groups.

In practice, comprehensive program evaluation usually involves multiple evaluation designs. The early stages of an evaluation may generate more ideas—answering an initial set of questions about a program often sparks additional questions. In these instances, a descriptive design may be most appropriate. On the other hand, the later stages of an evaluation often provide stronger

evidence of the impact of a program, as the questions of interest may be more refined. At this point in the evaluation, a QED or RCT may best address the evaluation questions.

What would such a comprehensive program evaluation look like? Let's say a district conducts a needs assessment to better understand how to support struggling students in math. Through the needs assessment, the district learns that some students are not prepared to complete their math homework individually. This is an example of a descriptive design. This district decides to offer additional support to students after hours at a single school, using AMMP!. The school where AMMP! is being implemented begins to see improvements in homework completion rates and overall math achievement. This is an example of a correlational design. To provide stronger evidence that the program is effective, the district conducts a pilot program with six middle schools within the district, using a lottery to determine three treatment schools and three comparison schools for the program. This is an RCT. After finding positive effects from the RCT, the district rolls out the framework to all schools that wish to participate. The agency then compares the results of all participating schools to the results of nonparticipating schools, using a matching technique. This is a QED.

Your choice of evaluation design also relates to whether your evaluation questions are process or outcome focused. (See chapter 1 in Module 2 for more information on process and outcome evaluation questions.)

Process evaluation questions are useful for describing a program and its implementation. Example process evaluation questions are "Is the program being implemented as intended?" "Are the program activities being conducted according to schedule?" and "What needs to be improved in the program, and how?"

Outcome evaluation questions, on the other hand, relate more to the results of a program. Example outcome evaluation questions are "What are the effects of the program?" "How can the program be sustained or replicated?" and "Were the intended outcomes achieved?"

In general, you can use descriptive and correlational designs to answer both process and outcome evaluation questions. Outcome evaluation questions that you answer using descriptive and correlational designs cannot provide evidence of efficacy. On the other hand, you can use QEDs and RCTs to answer outcome evaluation questions, which can provide evidence of efficacy.

For instance, in the prior AMMP! example, the evaluation included a descriptive design, a correlational design, a QED, and an RCT. The descriptive and correlational designs provided process information on student needs and informed the type of supports that might be helpful. The QED and RCT yielded outcome information, providing insight into how effective the program is and whether the outcomes have been achieved.

Next, in chapter 2, you will learn about threats to validity in evaluation design.

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